



# **an introduction to physics applications**

**Editor-in-Chief**

**Professor Pradip Narayan Ghosh**  
**University of Calcutta**



**UNIVERSITY OF CALCUTTA**



**An Introduction to Physics Applications Edited by Prof. P. N. Ghosh**

530  
In8

© University of Calcutta

G 16983

T3CU 2781

**Price : Rs. 150/-**

**Printed and Published by Pradip Kumar Ghosh,  
Superintendent Calcutta University Press,  
48, Hazra Road, Calcutta-700 019**



## **Editorial Board**

Professor Manoranjan Saha, Calcutta University

Professor Dilip Banerjee Do

Professor Santosh Sarkar Do

Professor Tapan Kumar Das Do

Professor Jyotiprasad Banerjee Do

Professor Asit Kumar Dutta Do

Dr. Ramen Kar Do

Professor Dibyendu Chakraborty, B.K. Girls' College, Howrah

Dr. Dipak Ghosh, R. K. M. Vidya Mandir, Belur, Howrah

Professor Jyotirmoy Dutta, B.E.S. College, Calcutta

Professor Harish Sarkar, Vidyasagar College, Calcutta

Dr. Tapas Bose, Scottish Church College, Calcutta

## Authors

Chapter — 1, 2	Dr. T. R. Bose Scottish Church College
Chapter — 3	Sri S. P. Ganchaudhuri Director, WEBREDA
Chapter — 4	Prof. D. Banerjee University of Calcutta
Chapter — 5—9	Dr. R. K. Kar University of Calcutta
Chapter — 10, 11	Prof. S. K. Sarkar University of Calcutta
Chapter — 12	Prof. A. K. Dutta University of Calcutta
Chapter — 13	Prof. D. Chakraborty Bijaykrishna Girls' College, Howrah
Chapter — 14	Prof. T. K. Das University of Calcutta
Chapter — 15	Dr. T. K. Ghosh Maharaja Manindra Chandra College, Calcutta



## Contents

### **SECTION—I** **Mechanics and Thermodynamics**

<b>Chapter 1</b>	<b>Heat Engine</b>	<b>1—31</b>
	1.1 Introduction	
	1.2 Heat Engine	
	(a) Broad Classification	
	(b) Essential features of the Heat Engines	
	(c) External and Internal combustion engine	
	(d) Principle of working of I.C. engine	
	(e) Air cycles	
	1.3 Internal Combustion Engines	
	(a) Different parts of the engine	
	(b) Operating cycle	
	(c) Valve timing	
	(d) Performance of Otto and Diesel Engine in relation to air cycle	
	(e) Detonation in Otto engine cycle, Engine knock	
	1.4 Indicated Horse Power and Break Horse Power	
	1.5 Rankine cycle	
	1.6 Questions and problems	
	1.7 References	
<b>Chapter 2</b>	<b>Energy Sources</b>	<b>32—78</b>
	2.1 Introduction	
	2.2 Fuels	
	2.3 Energy Storage	
	2.4 Turbine	
	2.5 Power sources—Power plants	
	2.6 Questions and Problems	
	2.7 References	





<b>Chapter 3</b>	<b>Non-conventional Energy Sources</b>	<b>79—88</b>
	3.1 Types of Energy	
	3.2 Solar Energy	
	3.3 Wind Energy	
	3.4 Biomass	
	3.5 Ocean Energy	
	3.6 Geothermal Energy	

<b>Chapter 4</b>	<b>Vacuum Techniques</b>	<b>89—103</b>
	4.1 Introduction	
	4.2 Qualitative description of the pumping process	
	4.3 Rotary oil pump	
	4.4 Mercury diffusion pump	
	4.5 Measurement of high vacuum	
	4.6 Leaks	

## **SECTION—II**

### **Electronics**

<b>Chapter 5</b>	<b>Feedback</b>	<b>104—111</b>
	5.1 Introduction	
	5.2 Basic Principles	
	5.3 Positive and Negative Feedback	
	5.4 Barkhausen Criterion	
	5.5 Oscillators	
	5.6 Questions and Problems	

<b>Chapter 6</b>	<b>Operational Amplifiers</b>	<b>112—129</b>
	6.1 Introduction	
	6.2 Characteristics	
	6.3 Concept of Virtual Ground	
	6.4 Uses of OPAMP	
	(a) Amplifier	
	(b) Mathematical Operation	
	(c) Filter	
	(d) Oscillator	





	6.5	A few glossary of OPAMPs	
	6.6	Questions and Problems	
Chapter 7		Miscellaneous Semiconductor Devices	130—138
	7.1	Introduction	
	7.2	Light Emitting Diode	
	7.3	Seven-segment Display	
	7.4	Silicon Controlled Rectifier	
	7.5	Diac and Triac	
Chapter 8		Digital Electronics	139—170
	8.1	Introduction	
	8.2	Combinational Circuits	
		(a) Adder and Subtractor	
		(b) Multiplexer and Demultiplexer	
		(c) Encoder	
		(d) Binary Coded Decimal (BCD) system	
		(e) Decoder	
		(f) More about 7-segment Display	
	8.3	Sequential Circuits	
		(a) Flip-flop	
		(b) J-K FF	
		(c) D-Flip-flop	
		(d) Master/Slave FF	
		(e) Edge triggering	
		(f) Registers	
		(g) Counters	
	8.4	Questions and Problems	
Chapter 9		Electronic Instruments	171—184
	9.1	Introduction	
	9.2	Cathode Ray Oscilloscope	
	9.3	Digital Multimeter	
	9.4	Measurements capacitance and Inductance	
		(a) Analog	
		(b) Digital	



### **SECTION—III**

#### **Communications**

<b>Chapter 10</b>	<b>Propagation of Electromagnetic Waves</b>	<b>185—208</b>
	10.1 Radio waves	
	10.2 Earth and atmosphere	
	10.3 Regions involved in radiowave propagation	
	10.4 Surface Wave Propagation	
	10.5 Space and propagation/tropospheric propagation	
	10.6 Sky wave propagation or ionospheric propagation	
	10.7 Exercises	
<b>Chapter 11</b>	<b>Transmission of Electromagnetic Wave</b>	<b>209—249</b>
	11.1 Modulation and Demodulation	
	11.2 Amplitude modulation	
	11.3 Frequency modulation	
	11.4 Detection or Demodulation	
	11.5 Detection of frequency modulated wave	
	11.6 Noise	
	11.7 Signal-to-noise ratio	
	11.8 Solved problems	
	11.9 Questions and Problems	
	11.10 Reference	
<b>Chapter 12</b>	<b>Signal Transmission Through Media</b>	<b>250—263</b>
	12.1 Introduction	
	12.2 Twisted pair and coaxial copper cable guided media	
	12.3 Optical fibre	
	12.4 Channel Capacity	
	12.5 Microwave communication	
	12.6 Satellite communication	
	12.7 Modems	
	12.8 Internet	





## SECTION—IV

### Computer

Chapter 13	Computer Hardware	264—283
	13.1 Introduction	
	13.2 Basic building blocks	
	13.3 Central processing unit	
	13.4 Storage system	
	13.5 Magnetic disks	
	13.6 CD-ROM	
	13.7 Input devices	
	13.8 Output devices	
	13.9 Software concept	
	13.10 Operating system concept	
	13.11 Computer programming	
Chapter 14	Programming in Fortran	284—298
	14.1 Introduction to Fortran	
	14.2 Constants and variable names	
	14.3 Arithmetic operators and modes for expressions	
	14.4 Fortran library functions	
	14.5 Mixed mode operation	
	14.6 Input output statements	
	14.7 Examples of simple programs	
	14.8 Control statements	
	14.9 The DO statements	
	14.10 Numerical methods	

## SECTION—V

### Experiments

• Chapter 15	Experiments 1-6	299—320
--------------	-----------------	---------

## **Preface**

The new three-year B.Sc. (General) course was introduced in 1998. The University authorities prescribed a guideline for this course. According to the guideline the syllabus should have an emphasis on applications. In order to understand the applications of physics in different fields we need elucidation of the basic laws of physics. The B.Sc. (General) Physics Part I syllabus is an attempt to introduce the basic laws and modern theoretical developments of physics. The third year syllabus is based entirely on the applications and techniques.

Physics deals with the fundamental questions of nature. When a branch of physics attains maturity so that its basic elements are comprehended as general principles, basic physics moves to applied physics or technology. Thus all of classical physics have resulted in applied physics, its contents formed the foundation of many branches of engineering. Discoveries of modern physics have rapidly been converted into technological innovations. Development of technology has always been in tune with the social needs. The large demand of energy prompted technologists to make use of the classical physics to develop power sources. The post-war developments of electronics and semiconductor devices have been extensively used for communication or transfer of information throughout the entire globe. The aim of the present book and the curriculum is neither description of basic laws of physics nor presentation of details of engineering techniques. The book deals with the application of physics principles in some of the technological fields. The technological applications of physics are vast and diverse in nature and span over almost all branches of engineering. This book presents introductory ideas on energy sources, electronics, communications and computer. A large number of books and reading materials are available on these subjects. Nevertheless the need for preparing this book arose from the fact that all the available books are usually in very detailed form and that one would need to go through too many books for the third year B. Sc. (General) course.

The entire material of this book is written and edited by experienced college and university teachers and experts in the field. The book is based on the University syllabus, but no guideline is given to the authors for preparation of the materials. Considering the diversity of the subjects it is difficult to have uniformity in the presentation, description and elaboration of the subjects. So the authors used their own freedom and judgement and for the sake of completeness they have included some topics which are not explicitly mentioned in the syllabus. But this may be useful to the teachers for acquiring a broad background knowledge of the topics discussed. This book must not be treated as a lecture note that can be presented to the students in its entirety.





In the section on computer programming we have included a brief idea of Fortran language only. The students may consult various other books available in the market. We have also given suggestions for books on the language C if any college prefers to teach the same instead of Fortran. It was not possible in the short time to prepare manuscript on both languages. This may be done in a future edition. The experiments to be performed are given in some details after actually working out them in the laboratory. Computer practicals are to be chosen by the teachers as per theory syllabus. Some chapters include questions and problems. But this could not be done for all the chapters.

After receiving the edited parts of the manuscripts from the members of the Editorial Board I have gone through the entire material and attempted to put things together in a coherent way as far as practicable. I thank all the authors and editors for the hard work they have put into it in spite of their busy schedules. I specially thank Sri S. P. Ganchaudhuri, Director, West Bengal Renewable Energy Development Authority and Prof. Subimal Sen, Secretary, Dept. of Science and Technology, Govt. of West Bengal for their help. We had only three months for preparation of the manuscripts and one month for editing. This may have caused errors and omissions. There may also be some repetitions. All these may be corrected in the next edition. I shall be glad if the teachers point out any such correction that they consider necessary. It should be mentioned that all the authors and editors acted on a completely voluntary and honorary basis. I thank Sri Prasanta Nandi of Physics Department of the University who helped in carrying out the experiments described in this book. I thank all the members of the Undergraduate Board of Studies and the teachers of different colleges who helped in preparation of the book. I thank Prof. Anil Bhattacharya for his advice in preparation of the manuscript. I am grateful to the former Vice-Chancellor of the University Professor R. N. Basu, the Vice-Chancellor Professor A. K. Banerjee, the Pro-Vice-Chancellor (Academic) Professor Surabhi Banerjee and Pro-Vice-Chancellor (Finance) Professor H. K. Banerjee for their kind help and interest. I thank Sri Sarajit Mullik for his help in preparing the diagrams. I also thank Sri O. S. Adhikary, Secretary, U. G. Council of Studies for his cooperation. I also thank Sri P. K. Ghosh, Superintendent and the staff members of Calcutta University Press for their keen interest and hard work in printing the book.

August 2000

Dean of the Faculty of Sciences  
University of Calcutta  
Department of Physics

Professor Pradip Narayan Ghosh  
Editor-in-Chief





## SECTION-I

# Mechanics and Thermodynamics

## Chapter 1

### Heat Engines

#### 1.1 Introduction

Civilization is a dynamic process. The process needs human activity. Activity means doing many things in many places. With the progress of civilization we need to do many more things in many more places.

These processes, however, need energy transfer, not merely energy. Energy sources are plentiful and are in many forms (conventional energy sources or non-conventional energy sources). However stored energy, as it is, cannot be put to useful purpose. For the utilization of stored energy we need energy transfer rather than energy.

Energy transfers are of two forms, thermal energy transfer or heat and mechanical energy transfer or work. A distinction between these two forms of energy transfer will be given shortly. Major part of our activity needs mechanical energy in the form of work. A part of the activity needs thermal energy. However, stored energy can be made readily available through thermal energy transfer (heat), rather than mechanical energy transfer (work).

So, we need a machine that will first convert stored energy to thermal energy and then the available thermal energy to work. This composite system is the power plant. The second part of the job is performed by a machine called heat engine. Work, heat and radiation energy transfer between different systems.

Thermal energy flows from a body at higher temperature to a body at lower temperature. Heat is this energy in transit due to the temperature difference. Any macroscopic system consists of a large number of particles (atoms or molecules), which are in a state of random motion. Temperature of a body is a manifestation of this random motion. Due to heat transfer, motion and configuration of these particles change. This shows up as a change of temperature or change of phase. In contrast, due to work, some macroscopic co-ordinates of the system, e.g. volume change. Consider, for example, a certain mass of gas contained in a cylinder fitted with a frictionless piston. As the gas expands through a volume  $\Delta V$  against a constant pressure  $P$  (Fig. 1.1) work done by the gas is :



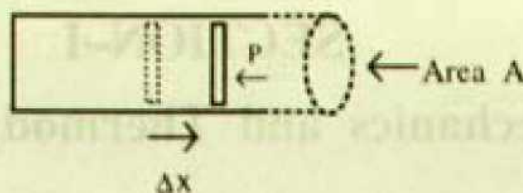


Fig. 1.1 Work done by expansion of gas

$$\Delta W = P \Delta V = PA\Delta X. \quad (1.1)$$

Note that the piston has moved through  $\Delta x$ , a change of external co-ordinate. Regarding energy source, chemical energy is essentially the electromagnetic energy of the atoms and molecules of the system. These energies are abundantly stored in the form of fossil fuels like coal, gasoline, diesel, etc. Such energies can easily be converted to thermal energy, resulting in increase of temperature. Heat energy is also available from windmills. Here the motive force, namely the wind force, originates due to pressure gradient, caused by unequal heating of the earth surface.

Thus, thermal energy is plentiful in nature. Thermal energy is directly utilized in many perposeful ways, like chemical processing, room heating, cooking, etc.

But, we use mechanical energy to a great extent for various useful purposes, for example in,

- 1) automobile, transport by water and aviation,
- 2) derivation of electrical energy from mechanical energy (Electrical energy can be transmitted over long distances via cables, only. There it can again be converted to mechanical energy for various uses),
- 3) machine works, process plants, etc.

In some cases mechanical energy is directly available; water power is one example. But, most of this energy is acquired by burning fossil fuel using a heat engine.

## 1.2 Heat engine

### a) Broad Classification

A heat engine is a device that absorbs heat from a system at high temperature, delivers useful work and returns some heat to a body at lower temperature. This device can continuously convert heat into work. Heat engines are broadly classified as

- 1) wind mills,
- 2) steam engine (external combustion engine),
- 3) internal combustion engine,
- 4) steam turbine, gas turbine.

The conversion of heat to mechanical work is brought about by some working substance. However, we need to produce continuous work and we want to reuse the

working substance. To achieve this, we need to circulate the working substance through a cycle of operations.

Such processes satisfy the following relation :

$$\oint dQ = \oint dU + \oint dW,$$

where  $dQ$  is the change of heat energy,  $dU$  is the change of internal energy and  $dW$  is the work done.

But,  $\oint dU = 0$  since  $U$  is a state function

$$\therefore \oint dQ = W$$

$W$  includes both available and wasteful work.

Since, the system operates cyclically, the working substance rejects some heat to the surrounding or sink at a lower temperature.

Let,  $Q_1$  = heat absorbed from the source at a higher temperature  $T_1$ ,

$Q_2$  = heat rejected to the sink at a lower temperature  $T_2$ .

Net heat absorbed :

$$Q = Q_1 - Q_2$$

We define efficiency of the engine as

$$\eta = \frac{\text{Work available}}{\text{Heat absorbed at higher temperature}}$$

If we assume that no work is wasted,

$$\begin{aligned} \eta &= \frac{W}{Q_1} \\ &= \frac{Q_1 - Q_2}{Q_1} \end{aligned} \quad (1.2.1)$$

Results of Carnot engine show that for such ideal engines operating reversibly, the maximum efficiency :

$$\eta_{rv} = \frac{T_1 - T_2}{T_1} \quad (1.2.2)$$

For all practical heat engines,

$$\eta < \frac{T_1 - T_2}{T_1}$$



b) *Essential features of Heat engines*

A heat engine requires the following basic units :

- 1) A source of heat of infinite heat capacity : It can supply large quantity of heat at a constant high temperature. Usually, coal, petrol, diesel, that store chemical energy, are burned to supply heat energy in the heat source.
- 2) A sink of infinite heat capacity : It can absorb large amount of heat discarded from the engine at a constant lower temperature. In a steam turbine the flames and the hot gases in the boiler form the hot reservoir. The cold water and the air used to condense and cool the used up steam constitute the cold reservoir.
- 3) A working chamber where the work is performed.
- 4) Working Substance : Exchange of heat with the source and the sink and transfer of work is brought about by the intermediacy of a substance called the working substance. Usually certain quantity of matter enters the engine, undergoes compression and expansion, undertakes the process of heat exchange with the surrounding and performs work.

In internal combustion engine, the working substance is a mixture of air and fuel; in steam turbine, it is water in different phases.

This is shown schematically in the energy flow diagram (Fig. 1.2).

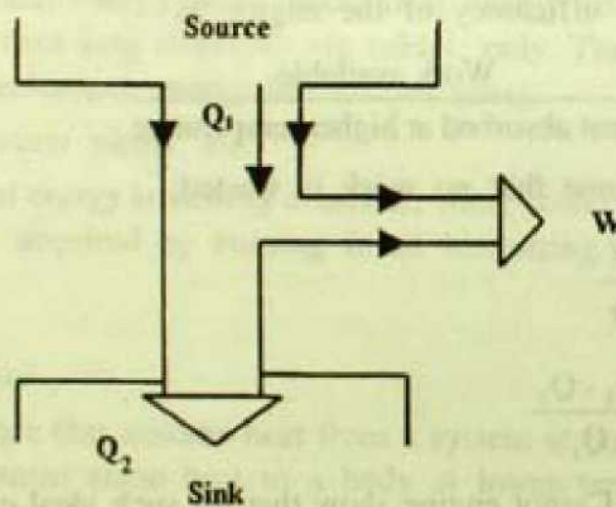


Fig. 1.2 Energy Flow diagram

Cross section of the flow tube is a measure of the amount of corresponding energy transfer. We get

$$Q_1 = W + Q_2$$

The working substance after performance of the work still has sufficient energy (unavailable at first instance) stored in it. In earlier versions of heat engine (e.g. steam engine), a condensing device was built in the working chamber. The used up steam was fed back directly by it to the boiler (the source). This caused a drastic



fall in efficiency. It was Sir James Watt who devised an external condenser with feed pump. The condenser performs work on the used up working substance, brings it to the initial state and feeds it back with some energy to the source. This greatly enhances the efficiency and helps to maintain a continuous supply of appreciable energy for a long time.

c) *External and Internal combustion engine*

In external combustion engine, the generation of heat from other sources of energy, outside the working chamber is carried out in a separate device, such as in a boiler. The working substance, a superheated vapour, carries the energy from the boiler to the working chamber and after the power stroke, it is fed back to the source by condenser and feed pump. The flow path for such engine is shown schematically in the figure (Fig. 1.3).

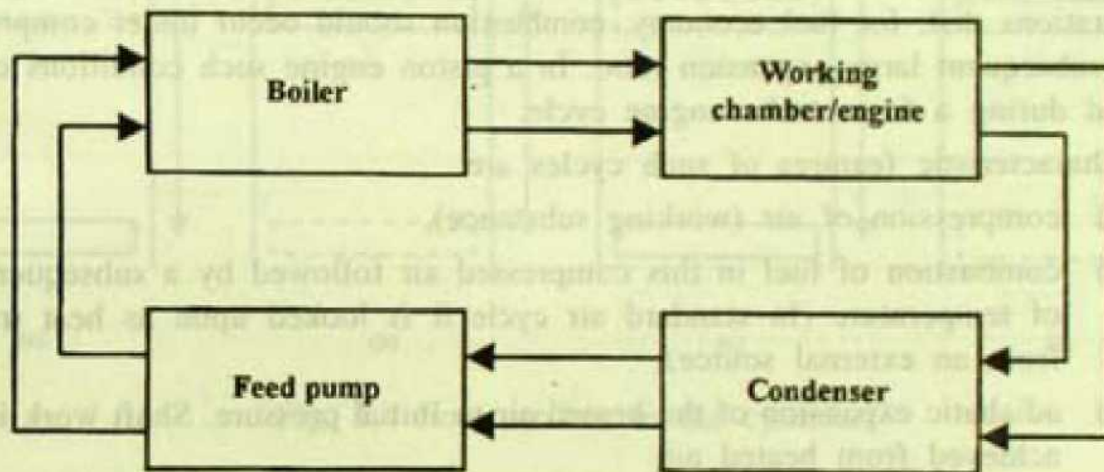


Fig. 1.3 Flow path

In internal combustion engine, the chemical energy of the fuel is released by combustion as heat energy, in the engine cylinder. Such engines do not use condensers. The residual heat is directly rejected to the surrounding, which acts as the sink. Supply of fuel air, the 'charge', is however continuous. There is loss of efficiency at the cost of perpetual available energy. In such engines one does not use the same air over and over. Here the engine operates cyclically, but the working substance does not pass through thermodynamic cycle. In this sense the internal combustion engine operates in an open cycle. However, for thermodynamic analysis, we treat them as closed cycle operations that approximate their actual operations. One such device is air-standard cycle. Such cycles only give us a tool to study qualitatively the performance of an engine in relation to different variables, like compression ratio, maximum pressure, etc. These engines are characterized by their high overall efficiency, light weight, compactness and low operating cost. These engines can be started and shut down as and when required.



To summarise, we can compare their performances as follows:

**External combustion engine**

- i) Fuel burnt in separate chamber.
- ii) Separate condensing device used.

**Internal combustion engine**

- Fuel burnt within the engine.
- No condensing device required.

Owing to the use of external source (furnace + boiler) external combustion engine becomes bulky. The ratio of power output to the weight is fairly low.

Internal combustion engines can be made much lighter. Power output/weight is fairly high. So internal combustion engines are more often used in light to heavy vehicle. External combustion engine, for the same reason, is limited to use in heavy land transport machines like, railway engine and stationary power plants.

d) *Principle of working of I.C. engine*

William Barnett (1838) and Beaudé Rocha (1862), showed from theoretical considerations that, for fuel economy, combustion should occur under compression with a subsequent large expansion ratio. In a piston engine such conditions can be achieved during a four stroke engine cycle.

Characteristic features of such cycles are

- 1) compression of air (working substance),
- 2) combustion of fuel in this compressed air followed by a subsequent rise of temperature (In standard air cycle it is looked upon as heat transfer from an external source),
- 3) adiabatic expansion of the heated air to initial pressure. Shaft work is thus achieved from heated air,
- 4) exhaust of the used up gas mixture. (In contrast to the intake and exhaust processes, the standard air cycle is completed by heat transfer to the surroundings.)

e) *Air Cycles*

***Otto Cycle***

The first engine to use this cycle successfully was built in 1876 by N.A. Otto.

In Otto engine, a mixture of fuel and air, 'the charge', is sucked into the cylinder. After desired compression, the charge is ignited by electric spark. So, Otto cycle is also called the spark-ignition combustion cycle.

For thermodynamic analysis in equivalent standard air cycle, we treat the air as the working substance. We further assume :—

- 1) The air behaves as an ideal gas. It obeys the perfect gas equation :

$$PV = n RT$$

- 2) The working substance does not undergo any chemical change. It merely acts as an agent to transport energy and perform shaft work.



- 3) Specific heats,  $C_p$  and  $C_v$ , remain constant throughout the process.
- 4) The cycles form an idealized reversible cyclic process. We ignore any friction present in the operation and also neglect turbulence and loss of heat to the cylinder walls.
- 5) We assume further that combustion is instantaneous.

Fig. 1.4 shows schematic diagrams for the motion of piston in the engine cylinder. This will explain the operations of Otto cycle. The displacement of the piston from the upper closed end to the other end is called a stroke.

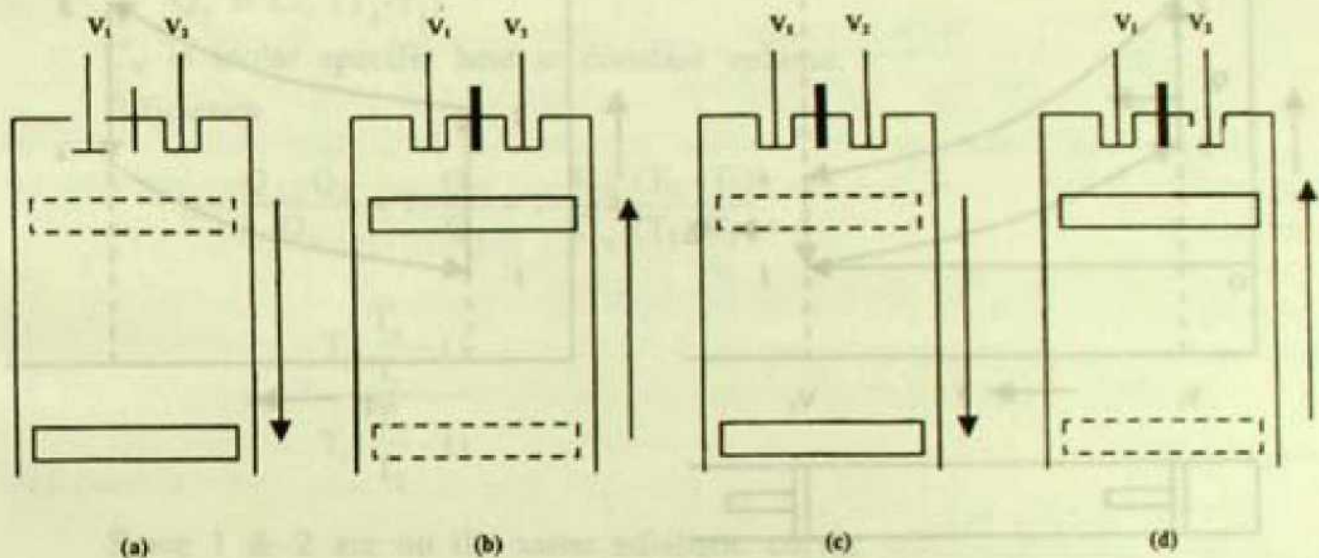


Fig. 1.4 Four strokes in the cylinder

a) *Isobaric suction stroke.*

The piston moves outward. Pressure inside the piston falls, the inlet valve  $V_1$  opens, the exit valve  $V_2$  remains closed. A proper mixture of air and petrol vapour (or gasoline) is sucked in isobarically by the outward motion of the piston.

b) *Compression stroke followed by ignition*

The air is compressed adiabatically to about  $1/8^{\text{th}}$  to  $1/10^{\text{th}}$  of its initial volume, by inward motion of the piston. The valves  $V_1$  and  $V_2$  both remain closed. Temperature and pressure both increase to a high value. Temperature is greater than the ignition point of the fuel.

A spark plug initiates a spark. The gas mixture ignites. The volume remains constant, temperature and pressure increase to a very high value.

c) *Working Stroke*

The valves,  $V_1$  and  $V_2$ , remain closed. The piston is thrown outward. The gas expands adiabatically to its initial volume. This is power stroke. The pressure falls to about atmospheric pressure.



d) *Exhaust Stroke*

The exhaust valve  $V_2$  opens. The piston moves inward; the used up gas mixture is discharged to the atmosphere. The pressure falls to atmospheric value.

The engine is ready for the next cycle of operations.

The operations are indicated in P-V diagram fig. 1.5(a). Entropy temperature diagram is shown fig. 1.5(b).

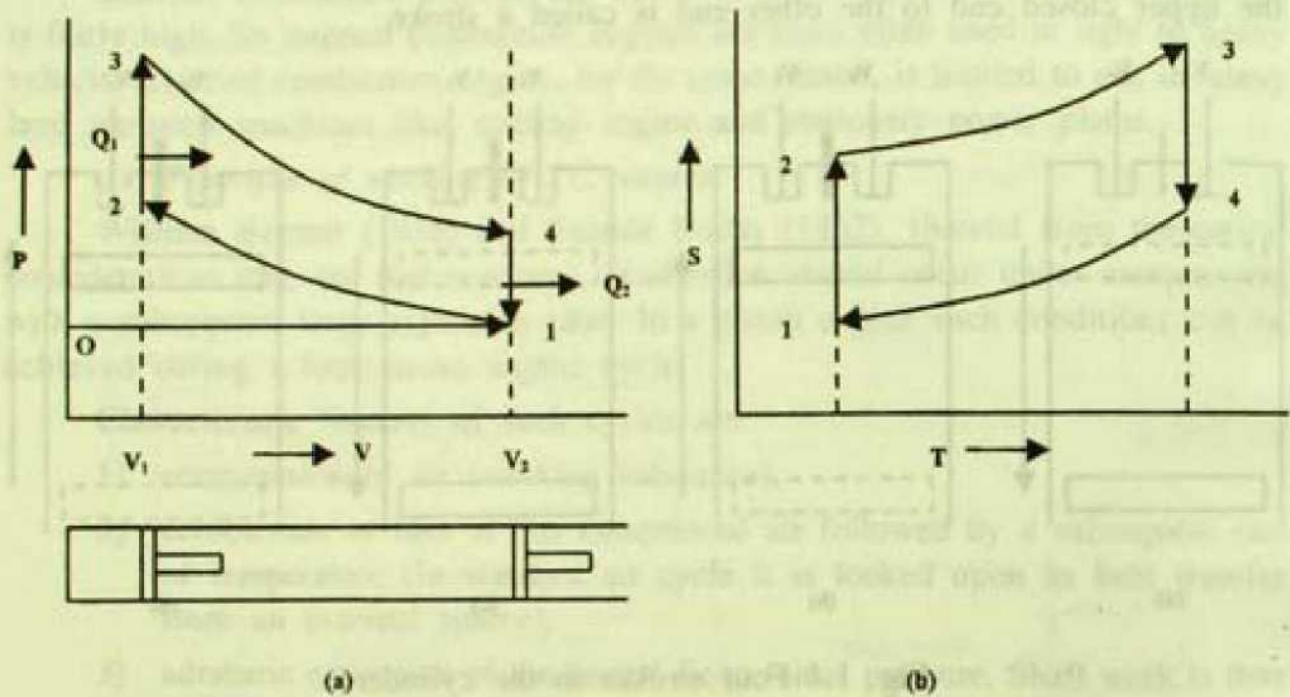


Fig. 1.5 (a) P-V diagram, (b) T-S diagram

0-1 : Isobaric suction stroke.

1-2 : Adiabatic compression. Both pressure and temperature increase.

2-3 : Isochoric ignition of the fuel.

At position '2', temperature is greater than ignition point of the mixture. Pressure also increases enormously. As the fuel mixture ignites, the piston motion is nearing the end of the upward stroke and finds no time to move back.

3-4 : Adiabatic expansion stroke of the gas the power stroke. At '4' the outlet valve opens.

4-1 : Isochoric scavenging stroke

The engine is ready for the next cycle of operations. We note that heat is both absorbed and rejected at constant volume of the working substance.

### Calculation of the efficiency of Otto Cycle

We assume, all processes are reversible.

Heat ( $Q_1$ ) is absorbed during the isochoric path 23 and it ( $Q_2$ ) is rejected during the isochoric path 4→1. Let  $T_1, T_2, T_3, T_4$  be the temperatures corresponding to the points 1, 2, 3, 4 respectively in the indicator diagram.

We consider 1gm-mole of the working substance

$$Q_1 = C_v (T_3 - T_2)$$

$$Q_2 = C_v (T_4 - T_1)$$

$C_v$  is molar specific heat at constant volume.

Efficiency

$$\eta = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} = 1 - \frac{C_v (T_4 - T_1)}{C_v (T_3 - T_2)}$$

$$= 1 - \frac{T_1 \left( \frac{T_4}{T_1} - 1 \right)}{T_2 \left( \frac{T_3}{T_2} - 1 \right)}$$

Since 1 & 2 are on the same adiabatic curve,

$$T_2 V_2^{\gamma-1} = T_1 V_1^{\gamma-1}$$

$$\alpha \quad \frac{T_2}{T_1} = \left( \frac{V_1}{V_2} \right)^{\gamma-1} = \rho^{\gamma-1} \quad (1.2.3)$$

where  $\rho = \frac{V_1}{V_2}$  is adiabatic expansion (or compression) ratio.

Again, 3 and 4 are on the same adiabatic curve.

$$T_4 V_2^{\gamma-1} = T_3 V_1^{\gamma-1}$$

$$\text{or} \quad \frac{T_4}{T_3} = \frac{1}{\rho^{\gamma-1}} \quad (1.2.4)$$

From eqs. 1.2.3 and 1.2.4

$$\frac{T_1}{T_2} = \frac{T_4}{T_3}$$



$$\frac{T_4}{T_1} = \frac{T_3}{T_2} = x \text{ and } \frac{T_1}{T_2} = \frac{1}{\rho^{\gamma-1}}$$

$$\eta = 1 - \frac{T_1 \left( \frac{T_4}{T_1} - 1 \right)}{T_2 \left( \frac{T_3}{T_2} - 1 \right)} = 1 - \frac{1}{\rho^{\gamma-1}}$$

We define mean effective pressure,

$$mep = \frac{\text{Work available}}{\text{Maximum volume change}}$$

It is, thus, defined as the constant pressure which if acted on the piston in its working stroke would generate the same amount of work as is actually done by the piston.

In Otto cycle, the mean effective pressure,

$$\begin{aligned} mep &= \frac{C_v (T_3 - T_2) - C_v (T_4 - T_1)}{V_1 - V_2} \\ &= \frac{C_v T_2 \left( \frac{T_3}{T_2} - 1 \right) - C_v T_1 \left( \frac{T_4}{T_1} - 1 \right)}{V_1 \left( 1 - \frac{1}{\rho} \right)} \\ &= \frac{C_v T_1 \left\{ \frac{T_2}{T_1} \left( \frac{T_3}{T_2} - 1 \right) - \left( \frac{T_4}{T_1} - 1 \right) \right\}}{V_1 \left( 1 - \frac{1}{\rho} \right)} \end{aligned}$$

For 1gm-mole of a diatomic gas,  $C_v = \frac{5}{2} R$

$$\begin{aligned} \therefore mep &= \frac{\frac{5}{2} R T_1 \left\{ \frac{T_2}{T_1} \left( \frac{T_3}{T_2} - 1 \right) - \left( \frac{T_4}{T_1} - 1 \right) \right\}}{V_1 \left( 1 - \frac{1}{\rho} \right)} \\ &= \frac{\frac{5}{2} P_1 \left\{ \frac{T_2}{T_1} (x - 1) - (x - 1) \right\}}{\left( 1 - \frac{1}{\rho} \right)} \end{aligned}$$



$$\therefore \text{mep} = \frac{\frac{5}{2} P_1 (x-1) \left( \frac{T_2}{T_1} - 1 \right)}{\left( 1 - \frac{1}{\rho} \right)}$$

### Characteristics :

Thermal efficiency depends on (i) the compression ratio and (ii) the ratio of specific heats,  $\frac{C_p}{C_v} = \gamma$ .

However, the compression ratio cannot be much increased.

At higher compression, temperature becomes quite high before the end of the stroke and the charge ignites before the sparking plug operates. This is known as pre-ignition.

### The Diesel Cycle (Diesel air cycle)

In Diesel Cycle, we attempt to increase the efficiency by increasing the compression ratio. To achieve this, fuel is introduced in the combustion chamber near the end of the compression stroke, not during suction stroke, as in Otto cycle.

The four strokes of the Diesel cycle are described below. The corresponding indicator curve is also explained. We refer to the Fig. 1.6.

- Suction stroke** : Inlet valve  $V_1$  is kept open, exit valve  $V_2$  remains closed. Air is sucked in isobarically by the outward motion of the piston. This is represented by the isobar ( $A \rightarrow 1$ ) in the P-V diagram (Fig. 1.7).
- Compression stroke** : The valves  $V_1$  and  $V_2$  are kept closed. The piston moves inward, compressing the gas adiabatically to about  $1/14^{\text{th}}$  of its initial value. Pressure and temperature increase. Temperature becomes greater than the ignition point of diesel. This is represented by the adiabatic curve ( $1 \rightarrow 2$ ) in the P-V diagram. (Fig. 1.7).
- Ignition (not a stroke)** : Near the end of the compression stroke, a third valve  $V_3$  opens, diesel is sprayed within the compressed air atmosphere by an inlet nozzle. The fuel ignites, chemical energy of liquid Diesel fuel is converted to heat energy. By proper control of piston motion, pressure is maintained constant. Volume increases, temperature also increases enormously. In the P-V diagram (Fig. 1.7) this is shown by the isobar  $2 \rightarrow 3$ .
- Working stroke** : Valves  $V_1$  &  $V_2$  remain closed near the state corresponding to '4', the valve  $V_3$  closes, the fuel supply is cut off. Thermal energy generated in the combustion chamber pushes the piston outward. The system undergoes the adiabatic expansion ( $3 \rightarrow 4$ ).

- e) Scavenging Stroke : Valves  $V_1$  and  $V_3$  remain closed. The exhaust valve  $V_2$  opens. The pressure in the used up gas falls to about atmospheric pressure. The piston moves inward, and pushes the used up fuel outside. This isobaric scavenging stroke is represented by the path  $4 \rightarrow 1$  in the P-V diagram. These processes complete the cycle of operations, the piston is now ready for the next cycle.

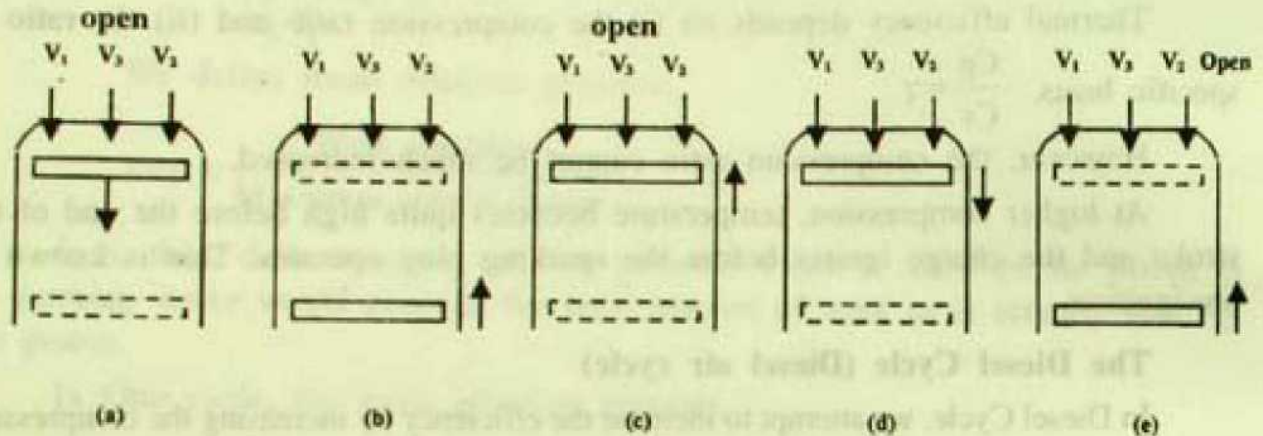


Fig. 1.6 Operations in a Diesel Cycle

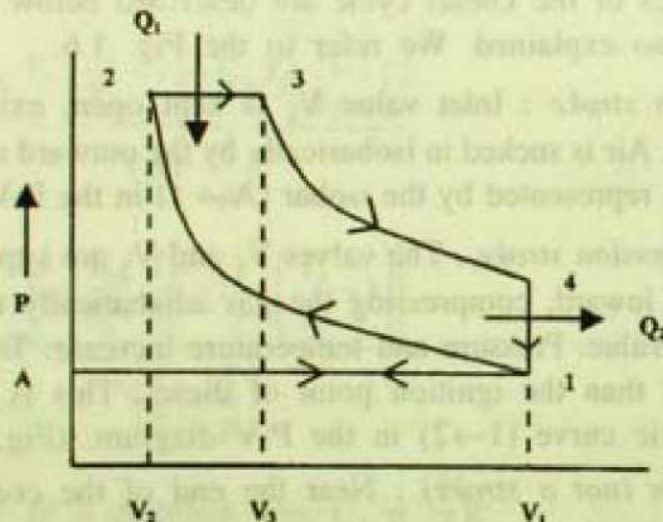


Fig. 1.7 P-V diagram

#### Calculation of efficiency :

We consider 1gm-mole of gas sucked in isobarically. We assume that it behave as a perfect gas. We note that here heat ( $Q_1$ ) is absorbed at a constant pressure and heat ( $Q_2$ ) is rejected at a constant volume of the working gas.

#### Efficiency of the cycle

$$\eta = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} = 1 - \frac{C_v(T_4 - T_1)}{C_p(T_3 - T_2)}$$



where  $C_p$ ,  $C_v$  are specific heat capacities at constant pressure and at constant volume respectively.  $T_1$ ,  $T_2$ ,  $T_3$ ,  $T_4$  are absolute temperatures corresponding to the point 1, 2, 3, 4 in the P-V diagram.

Efficiency of the Diesel cycle,

$$\eta = 1 - \frac{1}{\gamma} \frac{T_1 \left( \frac{T_4}{T_1} - 1 \right)}{T_2 \left( \frac{T_3}{T_2} - 1 \right)}$$

Since 3 and 4 lie on the same adiabetic,

$$T_4 V_1^{\gamma-1} = T_3 V_3^{\gamma-1} \Rightarrow T_4 = T_3 \left( \frac{V_3}{V_1} \right)^{\gamma-1} = T_3 \left( \frac{V_3}{V_2} \right)^{\gamma-1} \left( \frac{V_2}{V_1} \right)^{\gamma-1}$$

Since 2 and 3 lie on an isobar,

$$\frac{T_3}{T_2} = \frac{V_1}{V_2} = x$$

$$\therefore T_4 = T_2 \frac{x^\gamma}{\rho^{\gamma-1}}$$

Since 1 and 2 lie on the same adiabetic,

$$T_1 V_1^{\gamma-1} = T_2 V_2^{\gamma-1}$$

$$T_2 = T_1 \rho^{\gamma-1}$$

$$\therefore T_4 = T_2 x^\gamma$$

$$\therefore \eta = 1 - \frac{1}{\gamma} \frac{1}{\rho^{\gamma-1}} \frac{(x^\gamma - 1)}{(x - 1)}$$

Owing to higher values of compression ratio, this efficiency is slightly higher than the efficiency of Otto cycle.

**Comparison :**

To compare the different cycles we find the magnitude of efficiency of each of them and also the mean effective pressure and the maximum pressure the engine has to withstand.

We assume all the engines work between same temperature limits. In each case we take one gm mole of the gas as the working substance.

Otto cycle :

$$T_1 = 350 \text{ K}$$

$$T_2 = 620 \text{ K, ignition temperature of the fuel}$$

$$T_3 = 2100 \text{ K}$$

$$P_1 = \text{atmospheric pressure}$$

$$P_3 = \text{Maximum pressure}$$

$$\text{mep} = \text{Mean effective pressure}$$

$$= \frac{\text{Work available}}{\text{Maximum volume change}}$$

We write gas equation as

$$\frac{P_3 V_2}{T_3} = \frac{P_2 V_2}{T_2} \quad \text{or} \quad P_3 = P_2 \frac{T_3}{T_2} = P_2 x$$

$$\text{When } x = \frac{T_3}{T_2}$$

Since 1 & 2 lie on the same adiabetic,

$$T_1 V_1^{\gamma-1} = T_2 V_2^{\gamma-1} \Rightarrow P_1 V_1^{\gamma} = P_2 V_2^{\gamma} \Rightarrow P_2 = P_1 \left( \frac{V_1}{V_2} \right)^{\gamma} = P_1 \rho^{\gamma}$$

$$\therefore P_3 = P_1 \rho^{\gamma} x$$

$$\frac{T_2}{T_1} = \left( \frac{V_1}{V_2} \right)^{\gamma-1}$$

$$\rho^{\gamma-1} = \frac{620}{350}, \quad \gamma = 1.4, \text{ hence } \rho = 4.17$$

$$\text{It follows, } P_3 = P_1 \rho^{\gamma} x = 1 \times (4.17)^{1.4} \frac{2100}{620} = 25 \text{ atmos}$$

Mean effective pressure,

$$\text{mep} = \frac{\frac{5}{2} \times 1 \left( \frac{210}{62} - 1 \right) \left( \frac{62}{35} - 1 \right)}{1 - \frac{1}{4.17}} = 7.82 \text{ atmos.}$$



$$\text{Efficiency, } \eta = 1 - \frac{1}{\rho^{\gamma-1}} = 1 - \frac{T_1}{T_2} = 43.54\%$$

Diesel cycle

Considering fig. 1.7 for one gm-mole of a diatomic gas,

we take

$$T_1 = 350 \text{ K}$$

$$P_1 = 1 \text{ atmos}$$

$$T_2 = 930 \text{ K, burning point of the fuel}$$

$$T_3 = 2100 \text{ K}$$

$$P_3 = 35 \text{ atmos}$$

$$P_3 = \text{Maximum pressure}$$

$$\therefore x = \frac{T_3}{T_2} = \frac{2100}{930} = 2.258$$

$$\therefore T_4 = T_1 x = 350 \times 2.258 = 790.3$$

$$\therefore \rho^{\gamma-1} = \frac{T_2}{T_1} = \frac{93}{35} \Rightarrow \rho = 11.51$$

$$\eta = 1 - \frac{1}{\gamma} \left( \frac{1}{\rho^{\gamma-1}} - \frac{x^{\gamma}-1}{x-1} \right)$$

$$= 54\%$$

$$\therefore \text{mep} = \frac{Q_1 - Q_2}{V_1 - V_2} = \frac{C_p(T_3 - T_2) - C_v(T_4 - T_1)}{V_1 - V_2}$$

$$= \frac{C_p(T_3 - T_2) - C_v(T_4 - T_1)}{V_1 - V_2} = \frac{\frac{7}{2}R(T_3 - T_2) - \frac{5}{2}R(T_4 - T_1)}{\frac{RT_1}{P_1} \left( 1 - \frac{1}{\rho} \right)}$$

$$\therefore \text{mep} = \frac{\frac{7}{2}(2100 - 930) - \frac{5}{2}(790.3 - 350)}{350 \left( 1 - \frac{1}{11.51} \right)}$$

$$= 9.37 \text{ atmos}$$

The following table will help in comparing the performance of Otto and Diesel cycle.

	Otto cycle	Diesel cycle
$\rho =$	4.17	11.51
$P_c =$ (Maximum Pressure)	25 atmos	35 atmos
$mep =$	7.82 atmos	9.37 atmos
$\frac{P_c}{mep} =$	3.2	3.7
$\eta =$	43.6%	54%

The ratio  $\frac{P_c}{mep}$  measures the variations of stress inside the working chamber.

A lower value of this ratio is preferred. We note that Diesel engine should be more robust to withstand higher pressure and its variation. The efficiency of Diesel cycle is higher than the efficiency of Otto cycle. This is due to higher expansion ratio. For the same value of  $\rho$ , it can be shown that Otto cycle is more efficient than Diesel cycle.

### 1.3 Internal Combustion (I. C.) Engines :

#### a) *Different parts of the engine*

Figure 1.8 shows a sketch of the components of a four stroke internal combustion (I.C.) engine. Here we briefly describe the Diesel engine and occasionally refer to the Otto engine pointing out the essential differences between the two. The principal operating components are given. The fuel is burned within the cylinder (1) and power is developed. The cylinder head (2) at one end houses the inlet and outlet valves and the fuel supply nozzle in Diesel engine or the spark plug in the Otto engine.

The piston (3) closes the cylinder at the other end. The space bounded between the top closed end (also called the top dead centre or t.d.c.) and the other extreme position of the piston (bottom dead centre or b.d.c) forms the combustion chamber.

Power developed by the fuel, due to ignition, causes the piston to move and the linear to and fro motion of the piston is converted to the rotational motion of the fly wheel (4) via the piston rod (5), crank (6) and the crank shaft (7).

The flywheel is a disc of fairly large moment of inertia. The energy released



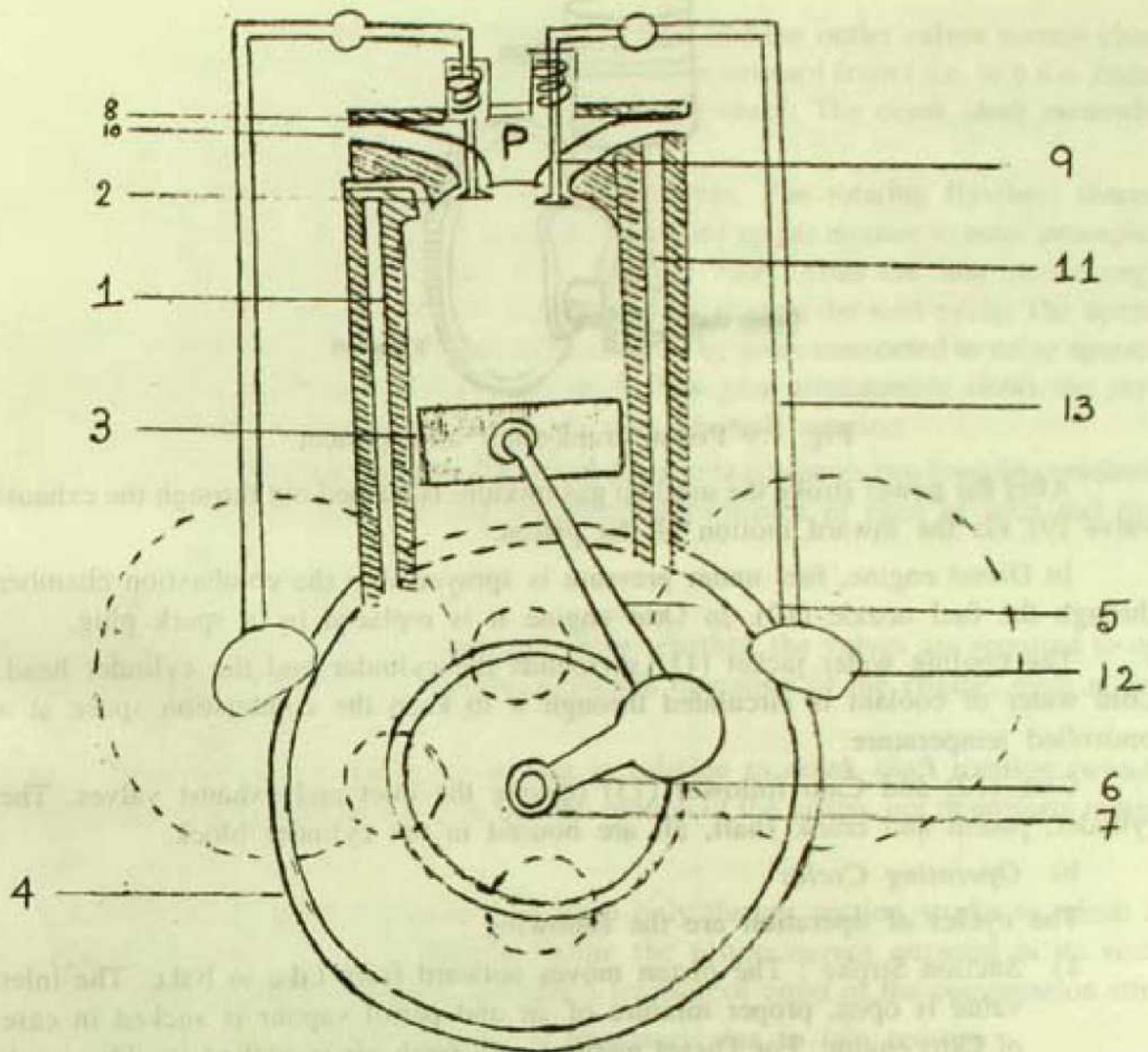


Fig. 1.8 Line diagram showing parts of an IC engine

during the power stroke is stored by the flywheel, and it helps to maintain the other strokes at the cost of a part of this energy. The balance of energy is available as the shaft work. The flywheel, owing to its high moment of inertia, maintains a fairly uniform output torque.

Fresh air in Diesel engine (or a proper mixture of air and petrol vapour in Otto engine) enters through the inlet valve (8) during the suction stroke.

On completion of the suction stroke, the inside pressure is still less than the atmospheric pressure and due to forward inertia the charge continues to pour in. The closing is determined as that position of the compression stroke at which the pressure of the charge equals one atmospheric pressure.

Optimum charging is achieved by opening the inlet valve  $15^{\circ}$  to  $20^{\circ}$  of the crank shaft rotation in advance of the t.d.c and closing it  $40^{\circ}$  to  $50^{\circ}$  of crank shaft rotation beyond the t.d.c. This is shown in figure 1.10. Exhaust valve timings

The exhaust valve opens about  $45^{\circ}$  of crank shaft rotation before the piston onsets its exhaust stroke from b.d.c. In this way the used up gas mixture, at a pressure higher than the atmospheric pressure, rushes out on its own. The exhaust valve is closed at about  $10^{\circ}$  of crank shaft rotation after the piston reaches the t.d.c. The incoming charge pressure further helps to push out the used up gas mixture. This is illustrated in the outlet valve timing diagram (Fig. 1.11).

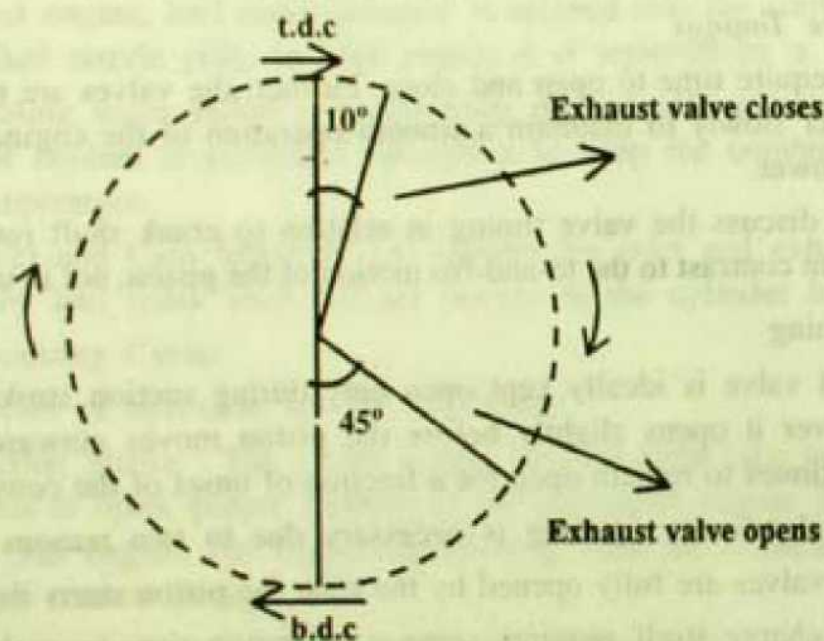


Fig. 1.11 Exhaust valve timing diagram

For a Diesel engine, a small time interval is required by the Diesel fuel to mix homogeneously with the compressed air before it ignites. This requires that the fuel be injected at about  $10^{\circ}$  to  $15^{\circ}$  advance of the piston arriving at the t.d.c. of the compression stroke.

#### Ignition advance (Otto engine)

Ignition should start when the piston, in its compression stroke, reaches the t.d.c. However, there is a time lag between initiation of spark and actual ignition of the fuel. So the spark is initiated in advance of the piston reaching the t.d.c. This is called ignition advance. However, a large ignition advance may cause a back explosion and turn the piston backward. This may cause the engine to run in the reverse direction.



### Curburrator (Otto engine)

In an Otto engine a mixture of air and fuel is charged in the cylinder during the suction stroke. The curburrator vaporizes the petrol and also atomizes it, i.e. smashes it into fine particles and mixes it with air.

### Air-fuel mixture

Usually the air-fuel mixture is around 15 : 1 with lower limits of 7:1 when the fuel is highly explosive. At upper limit of 20:1, the fuel burns irregularly. However, the optimum ratio depends on the speed of the engine.

#### d) Performance of the Otto and Diesel engine in relation to air cycles

It should be noted that the Otto cycle and the Diesel cycle discussed earlier, are just mathematical model cycles. Here, we assume all the idealized situations that are never achieved in practice. The performances of these cycles act as a standard for comparing the performance of actual Otto engine and Diesel engine. It should also be kept in mind that the air cycles are treated as closed cycles. Actual engine cycles are open cycles where the used up gas mixture is thrown out instead of being recycled. In actual engine, there are friction, turbulence, loss of heat to cylinder walls and many other effects. The cycle of operations is never reversible. All these combine and reduce the performance, i.e. the efficiency, of an actual cycle.

#### e) Indicator diagrams for four stroke air cycles and actual four stroke-engine cycle.

##### i) Otto engine

In fig. 1.12 (a) and (b) we show indicator diagrams for actual four stroke Otto cycle and four stroke Otto engine cycle.

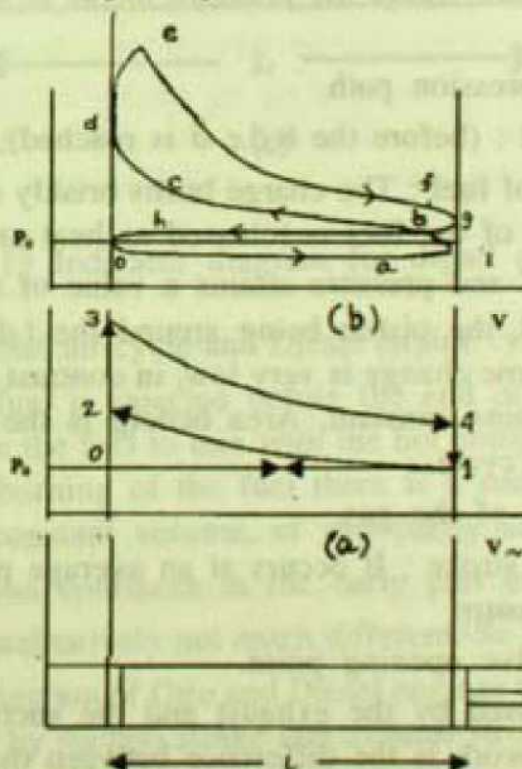


Fig. 1.12 Indicator diagram for Otto engine



They differ considerably, the reason for this difference is briefly summarised below with details, wherever felt necessary.

In ideal cycle, each of the four strokes is assumed to be operated by 180° crank rotation, and the valves should open and close only at t.d.c or b.d.c. as the case may be (fig. 1.12 (a)).

In actual operation for better performance these occur a few degrees on either side of the dead centres (discussed earlier in regard to valve timings). This is also shown in valve timing diagrams (figs. 1.10 and 1.11).

In the fig. 1.12, (a) is indicator diagram for actual Otto engine cycle and (b) represents indicator diagram for Otto air cycle.  $P_0$  is atmospheric pressure and  $L$  length of a stroke.

The valves are made to open and close slowly. This causes the rounding off of the corners of the indicator curve (fig. 1.12 (b)). As points of differences we note the following :

- 1) The suction and the exhaust lines are at slightly different pressures.
- 2) There is friction between the piston and the cylinder. There is considerable heat exchange between the working gas and the cylinder walls.
- 3) Products of combustion are heated only approximately at constant volume.
- 4) Specific heats of the gas mixture and the combustion products are different.
- 5) There may be dissociation due to chemical changes.

oa — Suction stroke : It occurs slightly below the atmospheric pressure.  
b — is the position where the pressure inside is equal to the atmospheric pressure.

bcd is the compression path

c — Spark point : (before the b.d.c d is reached); cd is ignition advance.

cde — Burning of fuel : The charge burns briskly at d. During combustion chemical energy of the fuel is released as heat energy. Temperature rises to about 2000°C and pressure attains a value of about 40kg/cm<sup>2</sup>. During this short period, the piston being around the t.d.c., its linear motion is rather slow, volume change is very low, in contrast to the ideal cycle where the volume remains constant, Area bdefg is the positive work done by the engine in a cycle.

ef — expansion of the gas.

fgh — exhaust stroke : It occurs at an average pressure greater than the atmospheric pressure

f — exhaust valve opening point.

The area aohb enclosed by the exhaust and the suction lines represents the pumpwork. The indicated work is the difference between the positive work and the pumping loss.



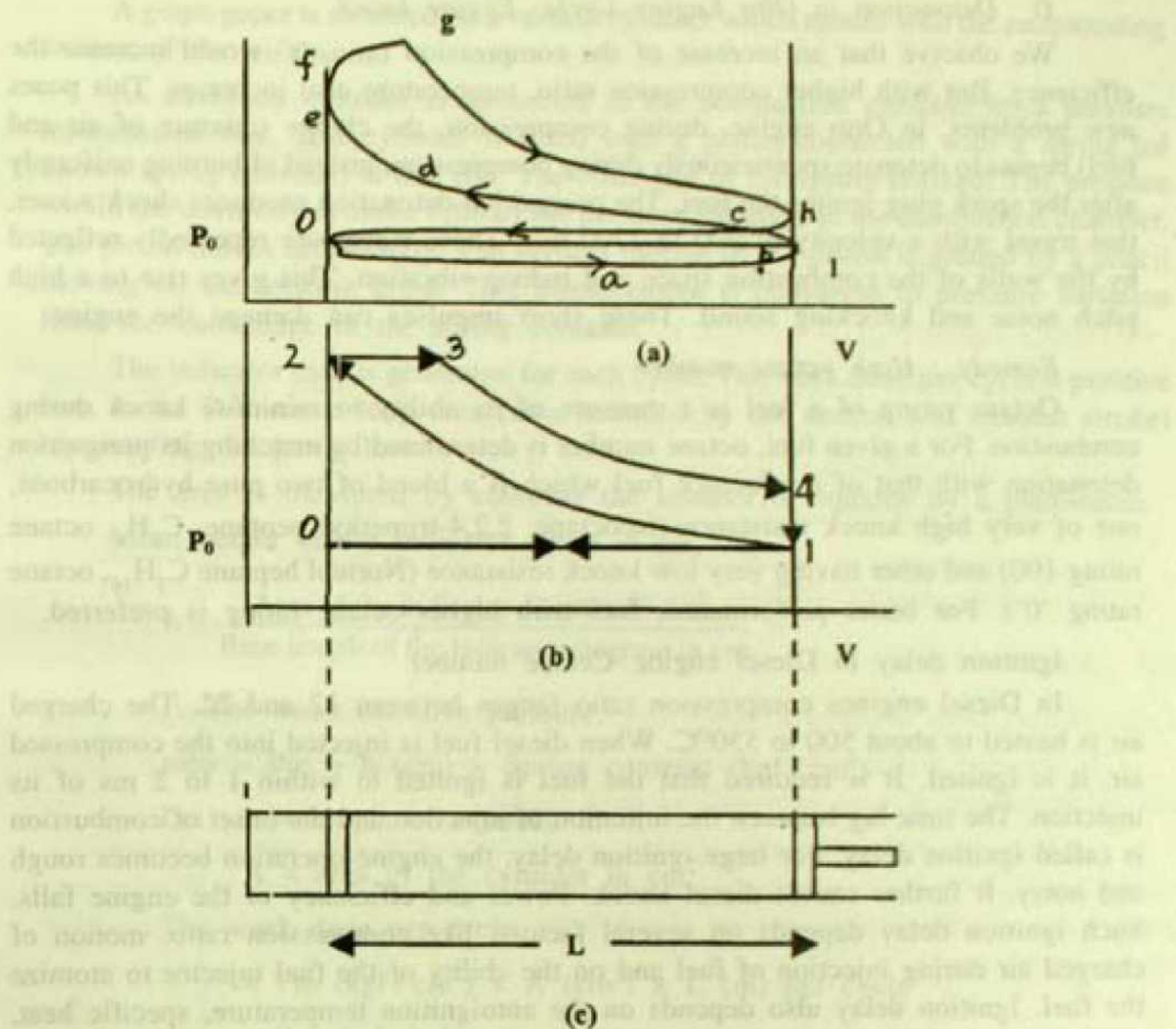


Fig. 1.13 Indicator diagram for diesel engine

Ideal four stroke Diesel air cycle and Diesel engine cycle are shown in fig. 1.13

In Diesel engines, fuel is injected before the end of the compression stroke. This gives enough time for the fuel to mix with the hot compressed air in the cylinder before it ignites. During burning of the fuel there is a sudden rise of pressure and combustion is partly at constant volume, ef and partly at constant pressure, fg.

Here also combustion continues in the early part of the expansion stroke.

Other features are qualitatively not much different. So we do not elaborate them.

In actual indicator diagram of Otto and Diesel engines the pump loss, represented by the closed area formed by suction stroke and exhaust stroke is difficult to measure, since it is rather small.



### f) *Detonation in Otto Engine Cycle; Engine knock*

We observe that an increase of the compression ratio, 'r' would increase the efficiency. But with higher compression ratio, temperature also increases. This poses new problems. In Otto engine, during compression, the charge (mixture of air and fuel) begins to detonate spontaneously during compression, instead of burning uniformly after the spark plug ignites the fuel. The preignition detonation produces shock waves, that travel with a velocity of 900 to 2700 m/s. These waves are repeatedly reflected by the walls of the combustion space and induce vibration. This gives rise to a high pitch noise and knocking sound. These short impulses can damage the engine.

#### *Remedy : High octane number*

Octane rating of a fuel is a measure of its ability to minimise knock during combustion. For a given fuel, octane number is determined by matching its preignition detonation with that of a reference fuel which is a blend of two pure hydrocarbons, one of very high knock resistance (isooctane, 2,2,4-trimethyl pentane,  $C_8H_{18}$ , octane rating 100) and other having very low knock resistance (Normal heptane  $C_7H_{16}$ , octane rating '0'). For better performance, fuel with higher octane rating is preferred.

#### *Ignition delay in Diesel engine: Cetane number*

In Diesel engines compression ratio ranges between 12 and 22. The charged air is heated to about 500 to 550°C. When diesel fuel is injected into the compressed air, it is ignited. It is required that the fuel is ignited to within 1 to 2 ms of its injection. The time lag between the initiation of injection and the onset of combustion is called ignition delay. For large ignition delay, the engine operation becomes rough and noisy. It further causes diesel knock. Power and efficiency of the engine falls. Such ignition delay depends on several factors, like compression ratio, motion of charged air during injection of fuel and on the ability of the fuel injector to atomize the fuel. Ignition delay also depends on the autoignition temperature, specific heat, density, thermal conductivity and surface temperature of the fuel.

#### *Cetane number*

It measures the ability of the Diesel fuel to ignite quickly after being injected within the cylinder. In laboratory, cetane number of a Diesel fuel is determined by matching its ignition delay with that of a reference fuel which is a blend of two fuels, one of short chemical ignition delay with a cetane number 100 and the other with large ignition delay, with a cetane number of zero.

### **1.4 Indicated Horse Power (I.H.P.) and Brake Horse Power (B.H.P.)**

Indicated Horse Power is the actual power developed within the engine cylinder. This is also equal to the work done per second by the combustion product minus the pump loss per sec.

The Indicated Power is measured by considering the indicator card produced by an attachment with the engine, called the indicator. In brief, the indicator is as follows :



A graph paper is mounted on a vertical cylinder which rotates with the reciprocating motion of the piston.

An auxilliary cylinder is connected to the combustion chamber by a pressure transmission line. This cylinder is fitted with a piston connected with a spring (of known spring constant) at one end. The other end of the spring is fixed. The pressure within the auxilliary cylinder follows the pressure variation in the combustion chamber. The piston moves accordingly. The vertical motion of the piston is plotted by a pencil moving on the rotating drum. This displacement is converted to pressure variation from the knowledge of the spring constant.

The indicator card is generated for each cycle. The work done per cycle = positive area in the P-V curve-loop area (area bounded by the suction and exhaust stroke) representing the pump loss.

The area is measured by counting the number of squares by a planimeter.

Mean height of the indicator curve in cm,

$$h = \frac{\text{Area of the indicator diagram in cm}^2}{\text{Base length of the indicator diagram in cm}}$$

∴ The mean effective pressure,

$$p_{me} = P_m = h \text{ (cm)} \times \text{Spring constant (kgf/cm}^2\text{)}$$

Let  $L$  = length of a stroke in m

$A$  = area of the cylinder in  $\text{cm}^2$

The work done per cycle

$$= P_m \text{ (kgf/cm}^2\text{)} \times A \text{ (cm}^2\text{)} \times L \text{ (m) per cycle}$$

$$= P_m A L \text{ kgf m}$$

In a four stroke engine, there is one cycle of operation for every two revolutions of the fly wheel.

Let  $N$  = revolution of the flywheel in rpm

∴ No of cycles of operation per sec.

$$= \frac{N}{2} \times \frac{1}{60}$$

Power indicated by the engine

$$= P_m A L \frac{N}{2} \frac{1}{60} \frac{\text{kgfm}}{\text{sec}}$$

$$1 \text{ H. P.} = 746 \frac{\text{Nm}}{\text{sec}}$$

$$= \frac{746}{9.8} \frac{\text{kgfm}}{\text{sec}}$$

∴ Indicated horse power

$$\begin{aligned} \text{I.H.P.} &= P_m A L \frac{N}{2} \frac{1}{60} \cdot \frac{9.8}{746} \\ &= \frac{P_m A L N}{2 \times 4567} \text{ H.P.} \end{aligned}$$

### Brake Horse Power

Brake power is the actual power available at the crank shaft. The brake power is less than the indicated power due to

- 1) resistance of the bearing,
- 2) air resistance on the motion of the flywheel and
- 3) power loss to operate other auxilliary units in tandem. Brake power is measured by a brake dynamometer.

## 1.5 Rankine Cycle

Cycle of operations in a steam engine under idealized conditions is the Rankine cycle.

It presents a thermodynamic cycle that acts as an ideal standard to compare the performances of heat engines and heat pumps.

In Rankine cycle the working substance is a condensable vapour (in contrast to idealized Carnot cycle, where the working substance is a perfect gas). In this respect, performance of a Rankine cycle more closely represents the performance of reciprocating steam engine cycle,

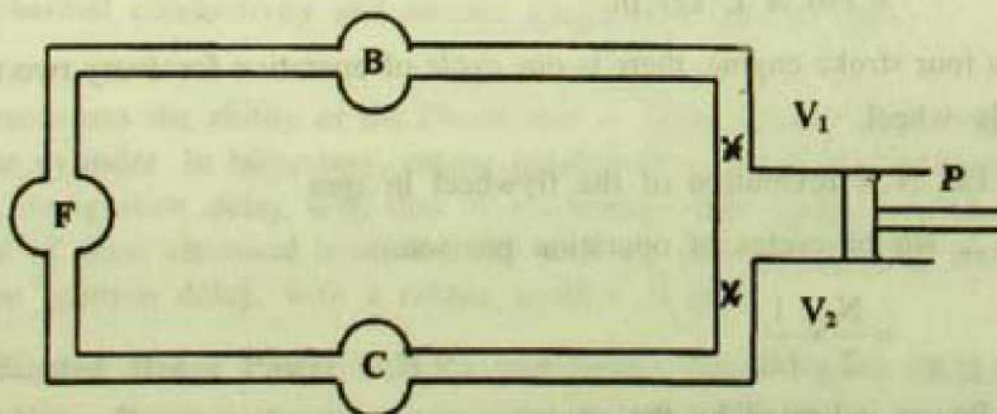


Fig. 1.14 Operation of Rankine cycle

Figure 1.14 represents a line diagram to represent the operations of the cycle with a condenser arrangement.



We make the following idealized assumptions :

- 1) All processes are reversible, quasistatic.
- 2) Irreversible processes like friction and loss of heat by conduction are absent.
- 3) Water behaves as a perfectly incompressible fluid.

Water from the condenser at temperature  $T_2$  and pressure  $P_2$  is forced by the feed water pump F to the boiler B at pressure  $P_1$  (Fig. 1.14). This isentropic, isochoric process is represented by AB (Fig. 1.15).

The temperature of the water rises to  $T_1$  at pressure  $P_1$  inside the boiler. The water is completely vaporised to steam and then to superheated steam. This is represented by the isobaric process BCDE (Fig. 1.15).

The steam now enters the engine cylinder via the inlet valve  $V_1$ . The steam pressure pushes the piston outward, and the pressure falls to  $P_2$ . This isentropic expansion is represented by EF (Fig. 1.15).

The used up steam is moved out through the exhaust valve  $V_2$  (Fig. 1.14) by the inward motion of the piston at constant pressure  $P_2$ .

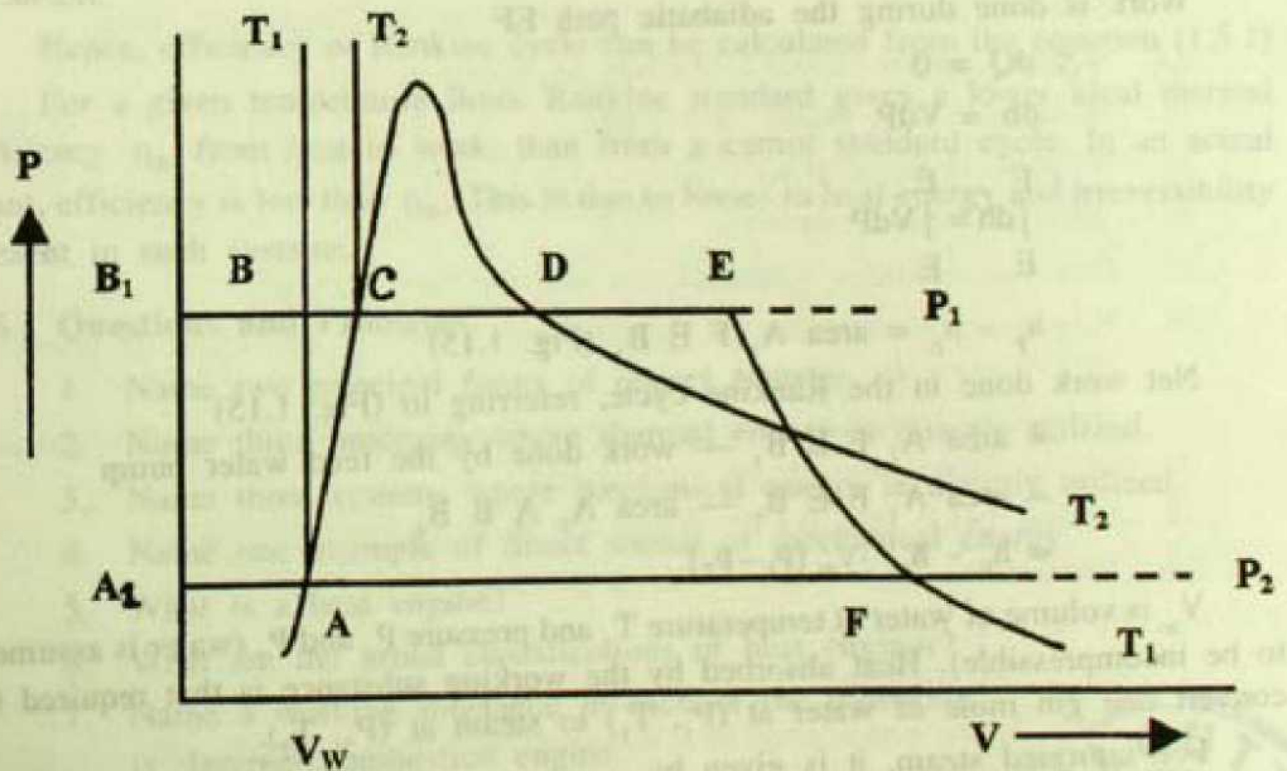


Fig. 1.15. Rankine cycle on P-V diagram

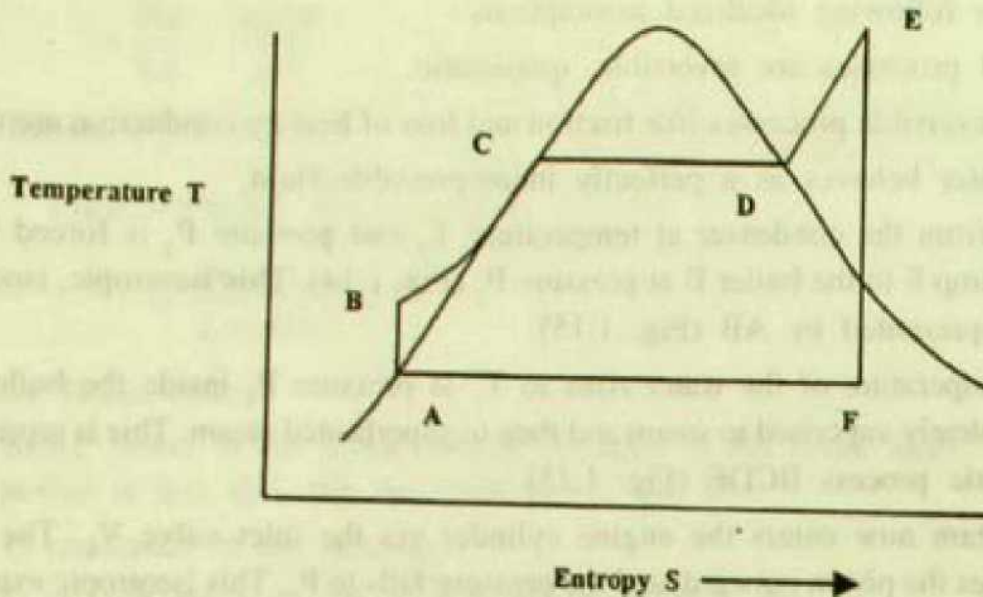


Fig. 1.16. Rankine cycle on T-S diagram

Calculation of Efficiency :

Consider 1gm-mole of water. Enthalpy function,

$$h = U + PV$$

$$dh = dU + PdV + VdP$$

$$= dQ + VdP$$

Work is done during the adiabatic path EF

$$\therefore dQ = 0$$

$$dh = VdP$$

$$\int_E^F dh = \int_E^F VdP$$

$$h_F - h_E = \text{area } A_1 F E B_1 \text{ (Fig. 1.15)}$$

Net work done in the Rankine cycle, referring to (Fig. 1.15)

$$= \text{area } A_1 F E B_1 - \text{work done by the feed water pump}$$

$$= \text{area } A_1 F E B_1 - \text{area } A_1 A B B_1$$

$$= h_F - h_E - V_w (P_1 - P_2)$$

$V_w$  is volume of water at temperature  $T_2$  and pressure  $P_1$  and  $P_2$  (water is assumed to be incompressible). Heat absorbed by the working substance is that required to convert one gm mole of water at  $(P_1, T_1)$  to steam at  $(P_2, T_2)$ .

For saturated steam, it is given by

$h_s - [h_w + V_w (P_1 - P_2)]$  where  $h_s$  is the total heat content of steam at  $(P_1, T_1)$  and  $h_w$  is the total heat of water at  $(P_2, T_2)$ .



∴ Efficiency

$$\eta = \frac{h_F - h_E - V_w (P_1 - P_2)}{h_s - h_w - V_w (P_1 - P_2)}$$

$$= \frac{h_F - h_E}{h_s - h_w} \quad (1.5.1)$$

The heat contents  $h_F$ ,  $h_E$ ,  $h_s$ ,  $h_w$  are obtained from the steam table.

*Mollier diagram (h-s diagram for different values of pressure p and dryness q).* Data on steam table are represented graphically in Mollier diagram (such graphs are standard and are available for estimation of h). There are two different families of h-s curves. One set of curves is drawn for different values of pressure p and temperature T. Other set of curves are drawn for different values of dryness, q. There are also boundary curves.

One boundary curve separates mixture of water and wet steam from dry, super heated steam.

The other boundary curve separates water from water and mixture of water and wet steam.

The critical isothermal is also shown in the diagram. Thus, from Mollier diagram one can estimate, under different conditions of temperature, the pressure and dryness of steam.

Hence, efficiency of Rankine cycle can be calculated from the equation (1.5.1)

For a given temperature limit, Rankine standard gives a lower ideal thermal efficiency  $\eta_R$  from heat to work, than from a carnot standard cycle. In an actual plant, efficiency is less than  $\eta_R$ . This is due to losses in heat energy and irreversibility present in such systems.

## 1.6 Questions and Problems

1. Name two principal forms of energy transfer.
2. Name three processes where thermal energy is directly utilized.
3. Name three systems where mechanical energy is directly utilized.
4. Name one example of direct source of mechanical energy.
5. What is a heat engine?
6. What are the broad classifications of heat engines?
7. Name a working substance in each of the following:
  - i) Internal combustion engine.
  - ii) Steam turbine
8. Give the energy flow diagram of a heat engine.





9. What are external and internal combustion engines?
10. Draw the flow path of an external combustion engine.
11. Describe in brief the principle of operation of an I.C. engine.
12. Describe in brief the operations of an Otto air cycle.
13. Calculate the efficiency of Otto cycle.
14. Define mean effective pressure.
15. Describe the operations of a Diesel air cycle. Draw its P-V indicator diagram. Calculate the efficiency of the Diesel cycle.
16. Show that for same compression ratio, the efficiency of an Otto cycle is greater than the efficiency of a Diesel cycle.
17. Describe in brief the different parts of a four-stroke I.C. engine.
18. Show that one complete I.C. cycle of a four-stroke engine causes two complete revolutions of the crank-shaft rotation and one complete performance of each of the inlet and the outlet valves.
19. Discuss the inlet valve timings in relations to the crank-shaft rotation.
20. Discuss the exhaust valve timings in relations to the crank-shaft rotation.
21. Discuss the differences between the Otto engine cycle and the Otto air cycle.
22. Discuss the nature of the indicator diagram for a Diesel engine cycle and compare it with the indicator diagram for the Diesel air cycle.
23. What is engine knock? Discuss its origin.
24. What is octane rating of a fuel?
25. What is ignition delay?
26. What is cetane number?
27. Can you consider windmills as heat engines? Explain.
28. How many revolutions of the flywheel results in one cycle of operation in a four-stroke engine?  
If there are 30 rpm of the flywheel, calculate the number of cycles of operations of the four-stroke engine per second.
29. Define I.H.P. and B.H.P. Which of them is lesser? Explain.
30. Explain the operations of the Rankine cycle in :  
i) a P-V diagram  
ii) a T-S diagram.
31. What is the importance of a Rankine cycle in relation to the operations of a power plant?
32. Find an expression for the efficiency of the Rankine cycle in terms of the enthalpy of steam and water.



## 1.7 References

1. A Treatise on Heat : M. N. Saha and B. N. Srivastava, The Indian Press (Publications) Pvt. Ltd., Allahabad
2. Heat Engineering : V. P. Vasandani and D. S. Kumar, Metropolitan Book Company Private Ltd., Netaji Subhas Marg, New Delhi.
3. College Physics : (Vol. 2) : D. B. Sinha and J. M. Das Sarma, Modern Book Agency, Calcutta.
4. University Physics : H. D. Young, R. A. Freedman, Addison-Wesley.
5. Thermal Engineering : P. L. Ballaney, Khanna Publishers.
6. Fundamentals of Classical Thermodynamics : J. Gordon and Richard E. Sonntag, Wiley Eastern Ltd.
7. McGraw Hill Encyclopaedia of Science and Technology.

## Chapter 2

### Energy Sources

#### 2.1 Introduction

We need energy to provide us heat, light and power that are necessary to sustain life and its activity. Every product, be it agricultural, industrial or service requires energy. In early days of human civilization energy sources were basically human and other animal power. To some extent wind power, solar power (to raise and dry crops) and water power were also used.

Non-conventional energy sources like solar energy, wind, water and tidal power are everlasting. They are available cost free. Their utilization is pollution-free. Such sources are known to us since long. But the cost of extracting useful energy from these sources remains quite prohibitive till date.

With the progress of civilization, in particular with the advent of industry, our requirement of power and energy is increasing day by day.

To name a few we require huge energy in industries like paper industry, primary metal, food products, chemical and allied products and for propulsion of ships, aircrafts, automotive power for trains, heavy trucks, cars etc. In cold countries a huge amount of energy is necessary for warming of houses alone. An estimated energy consumption in the year 2000 is around  $3.1 \text{ Q/year}$  for an estimated population of about  $10^{10}$  with an uncertainty factor of about 2.

$(1\text{Q} = 10^{18} \text{ BTu} = 1.7 \times 10^{11} \text{ bbl of petroleum equivalent, 1bbl} = 1 \text{ barrel} = 42 \text{ gallons})$

So for harnessing energy in order to meet the increasing demands of power we look forward to conventional energy sources.

These are

- 1) solid fuels,
- 2) liquid fuels,
- 3) gaseous fuels and  
(Liquid and gaseous fuels are mostly liquid hydrocarbons and gaseous hydrocarbons)
- 4) sources of nuclear energy.

The fuels are mostly fossil fuels and were formed millions of years ago. Non-fuel sources of energy include wastes, water, wind, geothermal deposits, biomass and



solar heat. Presently non-fuel sources contribute very little energy. But as fossil fuels, with their finite reserve, are used up, non-fuel source will become more and more important. These are renewable energy sources.

## 2.2 Fuels

### a) Solid fuels

The principal solid fuels are

- i) Coal : Anthracite, Bituminous coal, Subbituminous coal, Lignite or brown coal,
- ii) Peat and
- iii) Wood.

Composition of these fuels and calorific values are given in the table below

Table 2.1

Solid fuel	% analysis by weight						Calorific value k Cal / kg
	Carbon	Hydrogen	Oxygen	Nitrogen	Sulphur	Incombustible	
Anthracite	91.0	3.0	2.5	0.5	0.5	2.5	8500
Bituminous coal	81.0	5.0	8.0	1.5	1.0	3.5	7500
Lignite or brown coal	66.0	5.0	20.0	1.0	1.0	3.5	5000
Peat	58	6.3	30.8	0.9	—	4.0	3500
** Wood	48.5	6.0	43.5	0.5	—	1.5	2500

\* 1 kCal = 395.4 BTu

\*\* 1 Cord of wood =  $1.95 \times 10^7$  BTu; 1 Cord = 128 ft<sup>3</sup>)

A mid 1970 estimate of energy sources and their use is represented below in a  $\pi$ -Chart in Fig. 2.1.

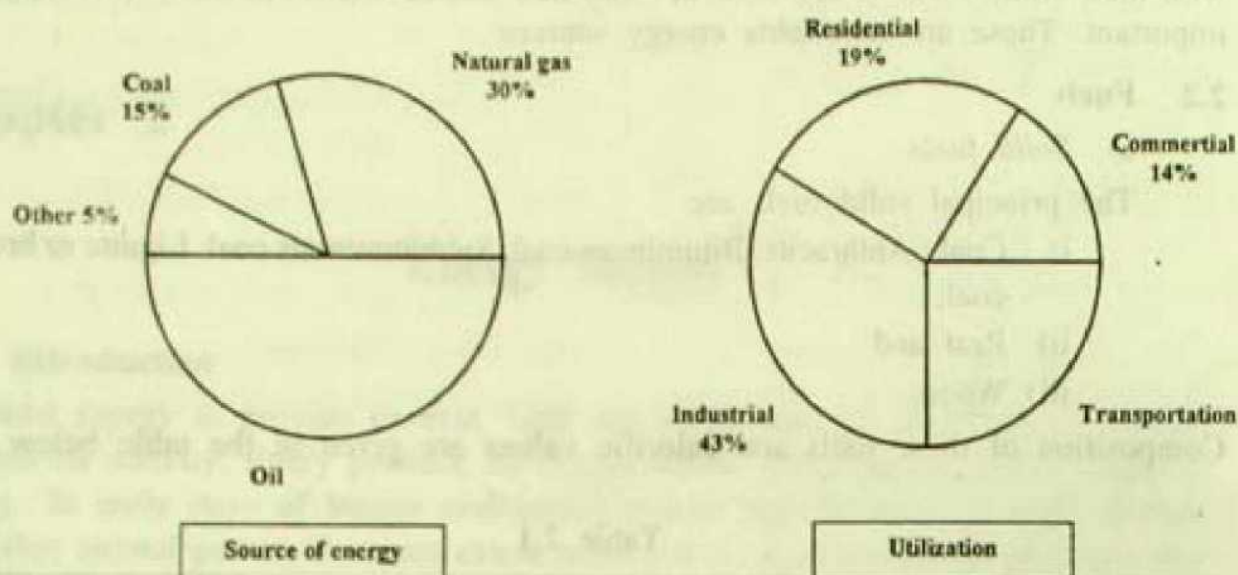


Fig. 2.1  $\pi$ -Chart for source and utilization of energy

Coal is formed from wood, vegetation and other organic matters 50 to 250 million years ago.

The best quality coal is Anthracite. It is hard coal. It is formed in older deposits and in deep beds. It has highest calorific value and it burns with little smoke.

Bituminous coal forms the major share of world coal reserve. It is extensively used for electric power generation, production of steel and coke. It is also used for house heating and has other industrial applications. Bituminous coal is also widely used in coal gasification plant to produce coal based gas, synthetic gas, principally, methane gas.

Brown coal contains about 50% by weight of water. Before use it is air dried so that it contains 10 to 20% of water. It is used as low grade fuel.

Annual production of coal in 1985 was around  $10^9$  Tons. Surface layers of coal are raised by method of area mining also called contour mining. Coal from underground are mined by method of continuous mining. This is done by method of room and pillar technique.

Mining of coal requires adequate manpower, proper transportation system and better technology.

Peat is formed from decay of plants and it contains about 90% of water. So it is to be air dried before use. In our country in villages and hilly terrains wood is still the most commonly used source of energy for cooking and house heating.



### b) *Liquid fuels*

Liquid and gaseous fuels are a mixture of many different hydrocarbons. Gasoline or petrol for example consists of a mixture about forty hydrocarbons. Some of the more important families of hydrocarbons are given below.

Table 2.2

Family	Formula	Structure	
Paraffin	$C_nH_{2n+2}$	Chain	Saturated
Olefin	$C_nH_{2n}$	Chain	Unsaturated
Diolefin	$C_nH_{2n-2}$	Chain	Unsaturated
Naphthene	$C_nH_{2n}$	Ring	Saturated
Aromatic Benzene	$C_nH_{2n-6}$	Ring	Unsaturated
Naphthalene	$C_nH_{2n-12}$	Ring	Unsaturated

A saturated hydrocarbon has all the carbon atoms joined by a single bond. In unsaturated hydrocarbon there are two or more adjacent carbon atoms joined by a double bond or a triple bond. Liquid fuels are primarily (i) gasoline or petrol, (ii) Diesel oil, (iii) crude oil or heavy fuel oil, (iv) paraffins.

Composition of these fuels and their calorific values are given in the table below

Table 2.3

Liquid fuel	% analysis by weight			Gross Calorific value kCal / kg
	Carbon	Hydrogen	Sulphur	
Gasoline or Petrol	85.4	14.6	—	11,200
Diesel oil	86.3	12.8	0.9	11,000
Heavy fuel oil/crude oil	86.1	11.8	2.1	10,500
Paraffin	86.3	13.6	0.1	11,100

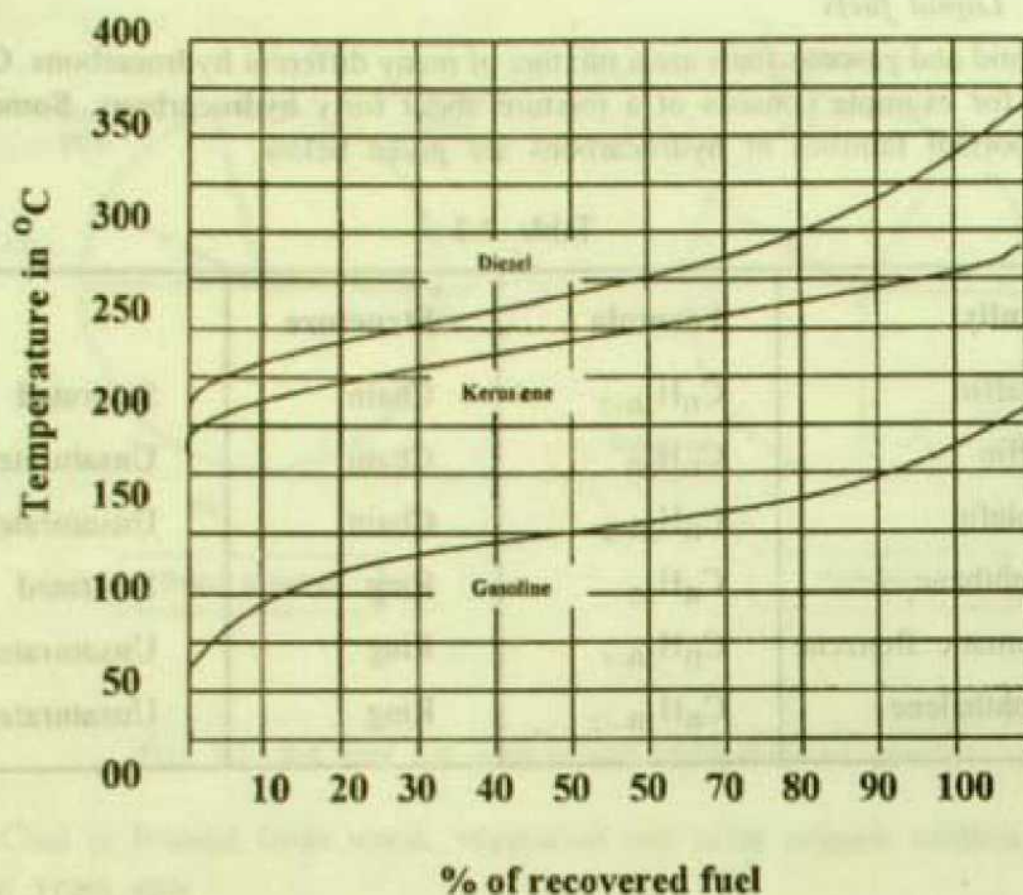


Fig. 2.2 Distillation curve

Alcohols are also used as fuels in internal combustion engines. Basically the liquid hydrocarbon fuels are mixtures of different hydrocarbons. From a given crude oil a variety of different fuels can be produced through distillation and cracking processes. Some of the more important ones are gasoline, kerosene, diesel oil and fuel oil and can be distinguished by the distillation curve: (Fig. 2.2). For this a sample of the fuel is slowly vaporized. The volatile hydrocarbon is distilled out first and condensed, others are distilled at still higher temperatures in the inverse order of volatility. The distillation curve Fig. 2.2 gives the fraction (%) of different fuels obtained as a function of temperature.

A given fuel is a mixture of many hydrocarbons, yet it is generally expressed in terms of a single hydrocarbon. e.g. gasoline/petrol is considered to be octane ( $C_8H_{18}$ ) and diesel fuel is considered as dodecane ( $C_{12}H_{26}$ ).

Liquid crude oil is obtained from underground, with the help of oil wells. It ranges from gasoline to very heavy viscous liquid with colors ranging from reddish green to black. Gasoline made from paraffin based petroleum contains hydrocarbons from  $C_5H_{12}$  (pentane) to  $C_{12}H_{26}$  (dodecane) and it boils between 30 and 220°C.



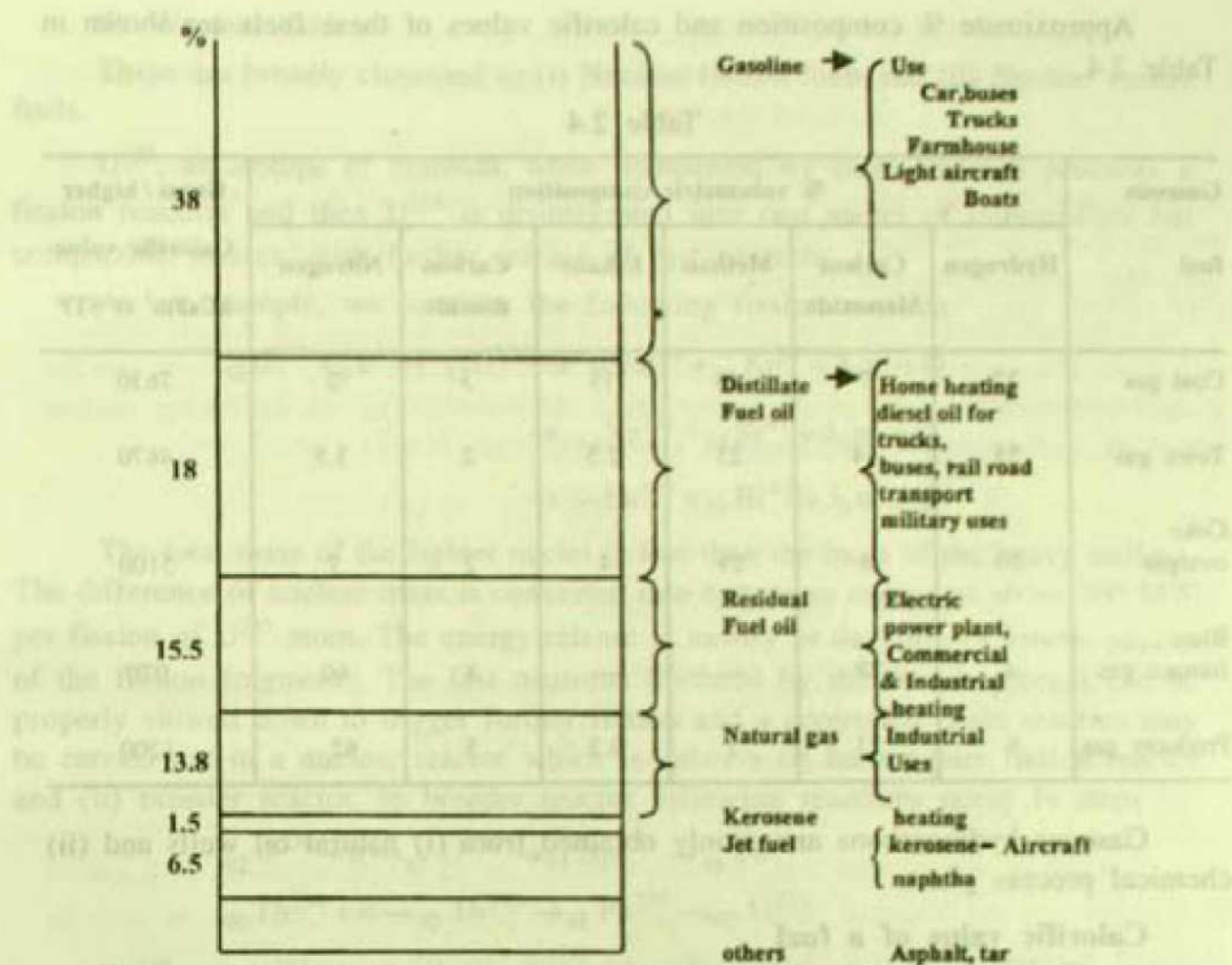


Fig. 2.3 Use of liquid fuels

Petroleum was formed in the sedimentary rocks million of years ago by the decay and incomplete oxidation of vegetation and animal remains. Oils may also be recovered from tar sands. These contain viscous hydrocarbon liquids in sand grain with silt and clay. Tar sands are the largest known source of liquid hydrocarbon. They are found throughout the world. Shale oil can be recovered from oil shale. It is a waxy solid hydrocarbon with clay, mud and silt. Refined oil is a major source of energy. To have an idea we show the different derivatives from a barrel of refined oil and their use in the fig. 2.3. Percentages (%) given are approximate.

#### c) Gaseous fuels

Some of the more commonly used gaseous fuels are (i) Coal gas, (ii) Town gas, (iii) Coke oven gas, (iv) Blast furnace gas, (v) Producer gas, (vi) synthetic gas like methane, (vii) Carbureted water gas, and (viii) various natural gases.

Approximate % composition and calorific values of these fuels are shown in Table 2.4.

Table 2.4

Gaseous fuel	% volumetric composition						Gross / higher Calorific value kCal/m <sup>3</sup> at STP
	Hydrogen	Carbon Monoxide	Methan	Ethane	Carbon dioxide	Nitrogen	
Coal gas	27	7	48	13	3	2	7630
Town gas	55	14	23	2.5	2	3.5	4670
Coke overgas	50	8	29	4	2	7	5100
Blast furnace gas	4	28	—	—	8	60	970
Producer gas	6	23	3	0.2	5	62	1200

Gaseous hydrocarbons are mainly obtained from (i) natural oil wells and (ii) chemical process plants.

#### Calorific value of a fuel

Calorific value of a fuel is defined as the heat energy liberated in kCal /kg for complete combustion of the fuel with adequate supply of oxygen. For gaseous fuels it is expressed in kCal / m<sup>3</sup> at STP.

Higher Calorific Value (H.C.V.) indicates total heat liberated in kCal / kg or in kCal / m<sup>3</sup> as the products of combustion, including the steam cooled to its initial temperature.

Lower Calorific value is the difference between higher calorific value and the heat absorbed by water produced therein to change it to vapour. An approximate calculation may be made as follows. We assume that the evaporation occurs at saturation temperature, corresponding to standard temperature of 15°C, latent heat of evaporation is 588.76 kCal / kg.

Lower Calorific Value (L.C.V.) is given by

$$\text{L.C.V.} = (\text{H.C.V.} - x \times 588.76) \text{ kCal / kg}$$

When  $x$  is fraction of water vapour due to combustion of 1 kg of the fuel.

There are elaborate methods for determining calorific value of a fuel, but we shall not go into those details.

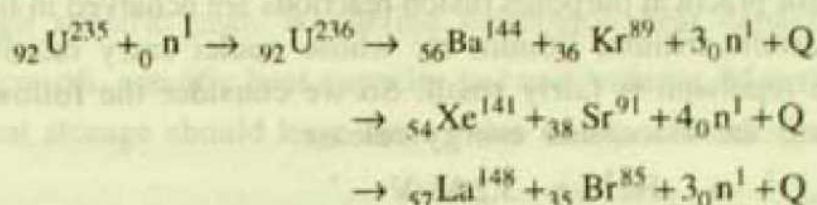


#### d) Nuclear fuels

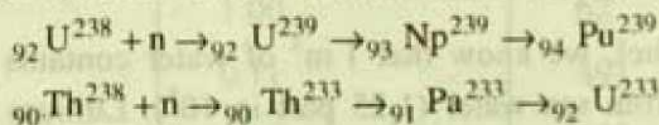
These are broadly classified as (i) Nuclear fission fuels and (ii) Nuclear fusion fuels.

$U^{235}$ , an isotope of uranium, when bombarded by slow neutrons produces a fission reaction and then  $U^{235}$  is disintegrated into two nuclei of intermediate but comparable masses, with further release of fast neutrons.

As an example, we consider the following fission reaction



The total mass of the lighter nuclei is less than the mass of the heavy nucleus. The difference of nuclear mass is converted into enormous energy of about 200 MW per fission of  $U^{235}$  atom. The energy release is mostly in the form of kinetic energy of the fission fragments. The fast neutrons liberated by the fission process can be properly slowed down to trigger further fission and a controlled chain reaction may be carried out in a nuclear reactor which is either a (i) burner pure fission reactor and (ii) breeder reactor. In breeder reactor following reactions occur in steps



$Pu^{239}$  and  $U^{233}$  are fissile products, thus formed by breeding  $U^{238}$ . These can be further used as fission fuels. Compared to a pure fission reactor the energy yield from a breeder reactor is more than 40 fold higher. Energy yield from 1kg of uranium is  $8 \times 10^{13} J$  which is equivalent to  $13.4 \times 10^3$  bbl of petroleum or  $2.8 \times 10^3$  ml coal or  $22.2 \times 10^3$  MWh or  $0.94 \times 10^3$  MW-d. To compare we consider an electric generation plant with an output of  $10^3$  MW. A fission reactor plant of this power output at a practical efficiency of 33% will require only 3.22 kg of  $U^{235}$  per day. However  $U^{235}$  is found as an isotope of uranium only at a concentration of about  $\frac{1}{140}$ . So a  $10^3$  MW plant requires about  $4.5 \times 10^2$  kg of uranium per day, which is equivalent to 42,400 bbl/d of petroleum. With a breeder reactor the energy output is increased many fold.

#### Nuclear fusion

In nuclear fusion two or more light nuclei combine to form heavier nucleus and energy is liberated by the process. In fact the binding energy per nucleon in the product nucleus is higher than that in the lighter nucleus. Consequently the total mass of the end nuclei is smaller than the sum of the masses of the combining nuclei and energy is released.

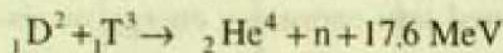
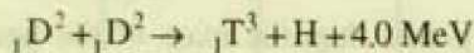
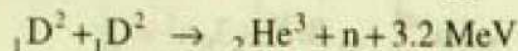


Energy released during fission

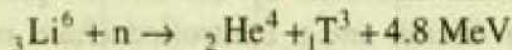
$$E = [\text{B.E. per nucleon of end nuclei} - \text{B.E. per nucleon of combining nuclei}] \times \text{no of nucleons}$$

Nuclear fusion reaction takes place under conditions of extreme temperature and pressure. This is necessary in order that the nuclei possess enough kinetic energy so that they can overcome their mutual electrostatic repulsion and can come closer than the range of nuclear forces.

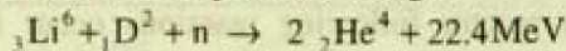
Consequently for practical purposes fusion reactions are achieved in light elements like hydrogen, deuterium, tritium, lithium etc whose nuclei carry relatively smaller charge and coulomb repulsion is fairly small. So we consider the following nuclear reactions and indicate the associated energy release.



When  ${}_3\text{Li}^6$  is bombarded by neutrons



From the last two equations we get



Regarding the nuclear fuel, we know that  $1 \text{ m}^3$  of water contains 34.4 gm of D. Lithium can be obtained from sea water (1.25 part in  $10^8$ ). Lithium can also be obtained from  $\text{Li}_2\text{O}$  which can be obtained from mining.

A rough estimation of the above data will show that from  $1 \text{ m}^3$  of water we can have energy release by deuteron fission of about  $5 \times 10^9 \text{ BTu} / \text{m}^3$  of water. To this information if we add that total amount of water available to us is  $\approx 10^{18} \text{ m}^3$ , the energy release in fusion process is around  $10^{10} \text{ Q}$  (yearly global energy consumption is expressed by  $1 \text{ Q} = 10^{18} \text{ BTu} = 1.7 \times 10^{11} \text{ bbl of petroleum equivalent}$ ).

### 2.3 Energy storage

There is still no practical and effective way to store mechanical and electrical energy output of a power plant. Often energy and power are abundantly available during periods of lean energy demand. Yet at peak hours of demand total available power output may not meet the demand. Power must be generated at the instant of its use. This puts a severe constraint on the functioning of power plants.

So we very much need to develop energy storage systems.

Energy storage requires means to collect and preserve readily available energy for future use. With an eye to achieve success in this area through extensive research,



here we very briefly describe the existing means of energy storage system.

### 1) Thermal Energy Sources

#### i) Sensible heat storage

Energy can be stored by heating a material to some elevated temperature

Heat stored this way  $Q = mS\theta = v\rho S\theta$

So for efficient energy storage the material should have high density and high specific heat. We define volumetric heat capacity.

$S_v = \rho S$ , specific heat capacity per unit volume. Material used for sensible heat storage should have large value of  $S_v$ . This is illustrated in Table 2.5.

Table 2.5

Suitable material for sensible heat storage

Material	Heat capacity BTu / lb / K	Density lb / ft <sup>3</sup>	Volumetric heat capacity
water	1.00	62	62
iron	0.11	491	54
magnesium	0.23	223	51
aluminium oxide	0.2	248	50
aluminium	0.12	168	37
glass	0.2	155	31
brick	0.24	125	30
concrete	0.21	140	29
dry earth surface	0.30	87	26

the table illustrates that water is a good medium of sensible latent heat storage 950 cuft of water at 100°F can store energy contained in 1bbl (24 gallon) of petroleum. Concrete walls help to store heat, rather reduce the temperature fluctuations due to diurnal heating.

ii) *Latent heat storage*

Energy is stored in the form of latent heat, by causing a phase change. There are no cost effective latent heat storage system near room temperature. Ferric chloride and lithium nitrate are useful as high temperature latent heat storage.

2) *Mechanical energy storage*

Energy stored in a rotating flywheel  $E = \frac{1}{2} I \omega^2$  where

$I \rightarrow$  moment of inertia of the flywheel about the axis of rotation.

$\omega \rightarrow$  angular velocity

Flywheels with large moment of inertia are used to store mechanical energy. Owing to its high rotational inertia, a flywheel helps to maintain a uniform supply of energy and power. Flywheel energy storage is widely used in all reciprocating engine systems.

In electric car an electric motor is used to store energy in a flywheel during stop ages of small duration. Between the stops a flywheel drives an electric motor as a generator that supplies part of the necessary power to run the car.

3) *Electrical energy storage*

Electrical energy may be stored as

- i) Capacitative storage,
- ii) Inductive storage and
- iii) Electric storage battery

The first two devices are used in electric power and electronic devices. At present large scale energy storage by these methods is not achieved. However large scale superconducting inductive energy storage system is a good possibility.

*Electric storage battery*

Such systems supply electrical energy from chemical energy. These are rechargeable batteries.

A storage battery designed for commercial use should have (1) high energy density, (2) low internal resistance, (3) durability, (4) fairly long life and (5) low production and maintenance cost.

Storage batteries are used in emergency lighting, emergency power supply, starting of automobile engines, railway carriage lighting, air conditioning and supplying power in submarines. It is widely used in portable television, radio



and in various military applications. Storage batteries are also used in telephone exchange. Such batteries are also used in the operation of photographic system.

Batteries are most reliable source of power. Many critical electric circuits are maintained and protected by battery power.

#### 4) *Pumped-Hydraulic storage*

In hydel project, during lean hours of demand, part of output energy is used to pump the water to a reservoir at considerable height. The head varies from about 250ft to 360ft. The stored energy is used at a suitable time to run a turbine and generates electrical energy.

#### 5) *Storage of liquid petroleum products*

There are both external and underground storage systems. It may be noted that (4) and (5) form the large scale storage systems.

## 2.4 Turbine

Turbines are the prime mover in majority of power plants. It will be more so in near future. So before we go over to power plants, in this section we study the operations of turbines in some detail.

Windmills that exist from antiquity are the precursors of present day turbines.

Wind blows due to unequal heating of earth. As the wind blows on the vanes of the windmills, it rotates and performs work. In this sense windmills fall in the category of heat engines. However windmills suffer from several drawbacks

- i) windmills can be installed only where wind force exists.
- ii) windforce is unreliable,
- iii) power output is rather small.

In turbines we generate artificial winds moving with high ordered velocity, that operates the turbine.

Turbines are broadly classified as

- 1) Steam turbines
- 2) Vapour or gas turbines

We discuss only steam turbines.

#### a) *Steam Turbines*

Here the energy of steam at high pressure and temperature is utilized to generate mechanical power in the form of rotational motion. The steam used in the turbines is produced in the boilers by burning fossil fuels like coal, oil, gas etc. To some



extent certain waste products are also used as fuels. To a large extent steam is also produced from the heat generated in nuclear reactors. Turbines are gradually replacing steam engines and I.C. engines for power generation. To understand its advantages and importance we first briefly note several inherent characteristics of the steam engines and I.C. engines.

In both cases (steam engines and I.C. engines), the working substance and the engine undergo a periodic change of temperature in a cycle of operation.

Also in such engines power developed is first converted to reciprocating motion of the piston, which is then converted to the rotational motion of the shaft, via crank and shaft arrangement. Thus the energy output is processed in two steps.

Both these combine to result in a loss in efficiency of the engine. Further, in such engines the thermodynamic energy of the working substance does not involve the external K.E. and external P.E. of the working substance. We may say 'the dynamic action of the working substance is not used.'

In contrast the very operation of steam turbine is based on the dynamic action of the working substance. Also the system together with the working fluid do not undergo periodic change in temperature.

Steam is first raised in a boiler by burning fossil fuel. This steam is under a very high pressure and temperature. It is then passed through a properly designed nozzle, which causes a pressure drop and thereby an increase in velocity. Thus certain amount of heat energy of the steam is converted to kinetic energy and the steam gets a high ordered velocity. This high velocity steam generates force and torque on the shaft either by impulse or by impulse together with reaction on the turbine. The shaft rotates. The turbine drives an alternator. The alternator converts mechanical energy into electrical energy. Accordingly steam turbines may be classified as (1) Impulse turbine and (2) Reactor turbine.

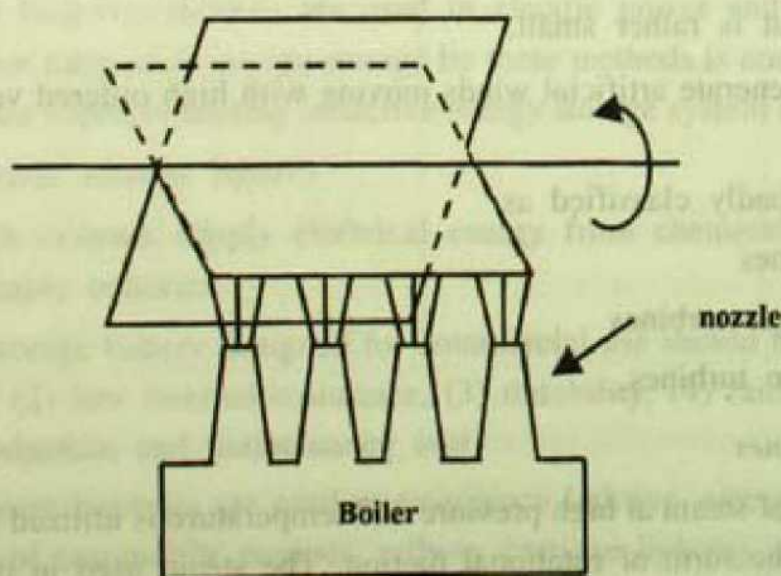


Fig 2.4 Impulse Turbine



Figure 2.4 is used to explain the principle of operation of impulse turbine.

### Step I

Steam from the boiler at high pressure and low velocity passes through a nozzle where pressure is dropped and velocity of steam increases.

### Step II

High velocity i.e. high momentum steam hits upon the turbine bladings. Direction of velocity changes. This causes an impulse. The impulsive force acts tangential to the turbine wheel and produces a torque about the shaft. The wheel turns.

#### *Reaction turbine*

Nozzles (bladings) are fixed to the periphery of the turbine wheel (fig. 2.5). Steam from the boiler enters the wheel and forces out tangentially with

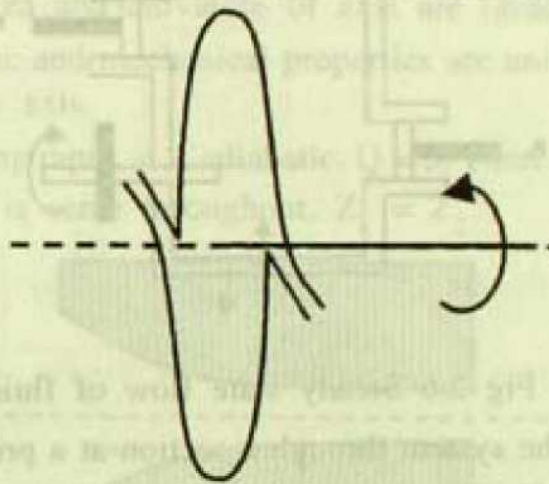


Fig 2.5 Reaction turbine

high velocity from the nozzles. Due to the reaction force the wheel turns in the opposite direction.

#### b) *Theory of steam turbine*

Theoretical analysis is done in several stages :— (1) Thermodynamics of flow process of high pressure steam; (2) function of the nozzle; (3) function of the turbine blades; (4) design of the turbine blades for maximum efficiency in relation to the incoming and outgoing steam jet.

##### 1) *Thermodynamics of flow process*

We consider a thermodynamic system and generalize the first law to include external kinetic energy, external potential energy and external work done by the system on a machine, called the shaft work. First law of thermodynamics for such a system can be written as

$$Q = \Delta U + \Delta K.E + \Delta P.E. + W \quad (2.4.1)$$

Where

$Q$  = heat absorbed

$\Delta U$  = increase in internal energy

$\Delta K.E.$  = increase in kinetic energy

$\Delta P.E.$  = increase in potential energy

$W$  = total work done by the system

We apply equation (2.4.1) to the steady state flow of a fluid through a device as indicated in the figure (2.6) below

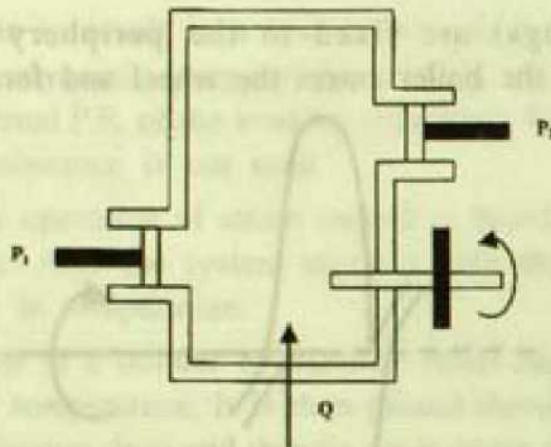


Fig 2.6 Steady state flow of fluid

The fluid enters the system through a section at a pressure  $P_1$ . With a velocity  $C_1$  at an elevation  $Z_1$ . The fluid leaves the system through a section at a pressure  $P_2$ , with a velocity  $C_2$  and at an elevation  $Z_2$ . To calculate the work done at entry and exit we imagine frictionless piston at these sections.

We consider unit mass of the fluid. Let  $U_1$ ,  $V_1$  and  $U_2$ ,  $V_2$  denote internal energy and volume per unit mass respectively at entry and at exit.

$W_s$  is the shaft work done per unit mass of the fluid.

$P_1 V_1$  = work done on unit mass of the fluid

$P_2 V_2$  = work done by unit mass of the fluid

Net work done by unit mass of fluid =  $(W_s + P_2 V_2 - P_1 V_1)$

$Q$  = heat absorbed

From equation (2.4.1)

$$Q = U_2 - U_1 + \frac{1}{2} (C_2^2 - C_1^2) + g (Z_2 - Z_1) + (P_2 V_2 - P_1 V_1 + W_s) \quad (2.4.2)$$

### Flow through a nozzle / Function of the nozzle

A nozzle is a duct of varying cross section. A pressure drop is maintained across it by some device. It is so designed that as the fluid (superheated steam or vapour



at high pressure) passes through it against the pressure gradient, it expands. As a consequence its velocity increases. The pressure drop along the nozzle effectively converts heat content of the fluid to external kinetic energy (to be shown later) and the fluid issues out with a very high ordered velocity. A proper design of the flow area together with the pressure helps in achieving this.

In developing equation (2.4.2) we tacitly assumed

- i) friction is absent,
- ii) state of the fluid at any point is the same at all times,
- iii) there is no change of chemical composition of the fluid; no change in chemical energy is involved.

For a nozzle we further assume

- iv) changes in area and curvature of axis are (gradual) small,
- vi) thermodynamic and mechanical properties are uniform over a cross section normal to the axis.

The fluid flow being rapid, it is adiabatic,  $Q = 0$ . There is no shaft-work involved  $W_s = 0$ ; the elevation is same throughout,  $Z_1 = Z_2$

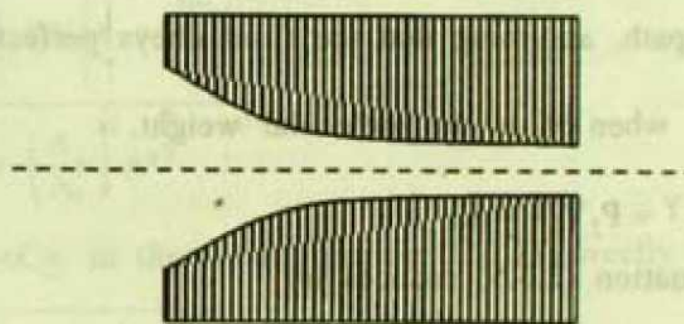


Fig. 2.7 Nozzle

∴ For isentropic flow through a nozzle, equation (2.4.2) for unit mass of fluid reduces to

$$0 = U_2 - U_1 + P_2 V_2 - P_1 V_1 + \frac{1}{2} (C_2^2 - C_1^2)$$

$$\frac{1}{2} (C_2^2 - C_1^2) = (U_1 + P_1 V_1) - (U_2 + P_2 V_2) \quad (2.4.3.)$$

$$\text{or, } \frac{1}{2} (C_2^2 - C_1^2) = h_1 - h_2 \Rightarrow \text{Heat drop} = \text{Gain in K.E.}$$

where  $h = U + PV$ , enthalpy per unit mass

Across any section of area  $A$  in the nozzle, mass of the fluid discharged per second.

$$m = \rho AC$$

$$= \frac{AC}{v}$$

$$v = \frac{1}{\rho} \rightarrow \text{Specific volume}$$

$$\frac{A}{m} = \frac{V}{C} \quad (2.4.4.)$$

This equation will be taken help of in designing the nozzle.

### Calculation of heat drop

We assume that the process is fast and adiabatic

Enthalpy per unit mass

$$h = U + PV$$

$$\begin{aligned} dh &= dU + PdV + VdP \\ &= dQ + VdP \end{aligned}$$

For isentropic (reversible adiabatic flow) processes

$$dQ = 0$$

$$\therefore dh = VdP$$

$$\int_1^2 dh = \int_{P_1}^{P_2} VdP \quad (2.4.5)$$

For isentropic path, assuming that the fluid obeys perfect gas law

$$PV = \frac{m}{M}RT \text{ when } M \text{ is the molecular weight.}$$

$$PV^\gamma = P_1V_1^\gamma = P_2V_2^\gamma = K$$

Substituting equation (2.3.5) reduces to

$$h_1 - h_2 = \frac{\gamma}{\gamma-1} P_1V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right) \quad (2.4.6)$$

$$\text{where we put } p^* = \frac{P_2}{P_1}$$

From equations (2.4.3) and (2.4.6)

$$C_2^2 - C_1^2 = 2(h_1 - h_2) = \frac{2\gamma}{\gamma-1} P_1V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right) \quad (2.4.7)$$

In the steady state from equation of continuity, mass of fluid per second through any section is same

$$\therefore \frac{A_1C_1}{V_1} = \frac{A_2C_2}{V_2}$$



$$\alpha \quad C_1 = \frac{A_2}{A_1} \cdot \frac{V_1}{V_2} C_2 \quad (2.4.8)$$

Substituting for  $C_1$  from equation (2.4.8) in (2.4.7)

$$C_2^2 \left[ 1 - \left( \frac{A_2}{A_1} \right)^2 \left( \frac{V_1}{V_2} \right)^2 \right] = \frac{2\gamma}{\gamma-1} P_1 V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right)$$

$$P_1 V_1^\gamma = P_2 V_2^\gamma$$

$$\left( \frac{V_1}{V_2} \right)^2 = \left( \frac{P_2}{P_1} \right)^{\frac{2}{\gamma}} = p^{*\frac{2}{\gamma}} \quad (2.4.9)$$

$$\therefore C_2 \left[ 1 - \left( \frac{A_2}{A_1} \right)^2 p^{*\frac{2}{\gamma}} \right] = \sqrt{\frac{2\gamma}{\gamma-1} P_1 V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right)}$$

$$C_2 = \frac{\sqrt{\frac{2\gamma}{\gamma-1} P_1 V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right)}}{1 - \left( \frac{A_2}{A_1} \right)^2 p^{*\frac{2}{\gamma}}} \quad (2.4.10a)$$

Usually  $C_1 \ll C_2$ , in that case equation (2.4.7) directly gives

$$C_2 = \sqrt{\frac{2\gamma}{\gamma-1} P_1 V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right)} \quad (2.4.10b)$$

Equation (2.4.10b) also follows from equation (2.4.10a)

From equation (2.4.9), specific volume of the fluid over any section

$$V = V_2 = \frac{V_1}{p^{*\frac{1}{\gamma}}} \quad (2.4.11)$$

For a given value of the pressure ratio  $p^*$  we can now calculate

- specific volume of steam at any section (equation 2.4.11)
- required area of cross section of the nozzle (equation 2.4.4) and
- speed of the steam (equation 2.4.10a or 10b).

We start with dry ( $q=0$ ) saturated steam at a pressure of  $15 \text{ kg/cm}^2$ . Table (2.4.1) gives velocity of steam jet and cross-section of the nozzle for different values of  $P_2$  i.e.  $p^*$

Table 2.4.1

Pressure $P_2 \text{ kg/cm}^2$	Dryness $q$	Volume in cubic metres	Velocity in m/s	Area of jet in $\text{cm}^2 \text{ sec/kg}$ $A/m = \frac{V}{C}$
15	1.	0.136	0	4.83
10	0.97	0.1933	400	4.77
9	0.963	0.2119	444	4.77
8	0.955	0.2347	492	4.87
7	0.946	0.2636	541	5.49
5	0.927	0.3528	644	5.49
1.033	0.856	1.430	957	14.94
0.12	0.784	9.85	1214	81.2

From an analysis of this data we note that with increase of pressure drop ( $p^*$  decreasing)  $A/m = \frac{V}{C}$  first decreases ( $V$  increases less rapidly than  $C$ ), next the area reaches a minimum, with further pressure drop, the area should increase and the issuing steam acquires a very high velocity.

Thus the nozzle should first converge, reaching a minimum area called the throat, thereafter the nozzle widens slowly in a diverging pathway, until the exhaust steam attains the desired high velocity for the preset drop of pressure. This is the theory of De Laval convergent-divergent nozzle.

Optimum area for maximum discharge per unit area is reached when

$\frac{A}{m}$  is minimum

$$\frac{A}{m} = \frac{v}{c} = \frac{V_1}{p^{*\gamma}} \cdot \frac{1}{\sqrt{\frac{2\gamma}{\gamma-1} P_1 V_1 \left( 1 - p^{*\frac{\gamma-1}{\gamma}} \right)}}$$

We put  $\frac{d}{dp^*} \left( \frac{A}{m} \right) = 0$  and get

$$p_c^* = \left( \frac{2}{\gamma+1} \right)^{\frac{\gamma}{\gamma-1}}$$

For  $p^* < p_c^*$  full velocity of the jet is attained with a convergent nozzle only.

To elaborate at this point let us make the following study that will help us to design a nozzle.



At any point in the nozzle equation (2.4.2) reduces to

$$Q = U - U_1 + PV - P_1V_1 + \frac{1}{2}(C^2 - C_1^2) + g(Z_2 - Z_1) + W_s$$

Put  $Z_2 = Z_1$

Differentiating the above equation we get

$$dQ = dU + PdV + VdP + d\left(\frac{1}{2}C^2\right) + dW_s \quad \dots(2.4.12)$$

$$dQ = VdP + d\left(\frac{1}{2}C^2\right) + dW_s$$

In a nozzle the shaft work  $W_s = 0$  at all points

$$\therefore 0 = VdP + d\left(\frac{1}{2}C^2\right)$$

$$\text{or } CdC + VdP = 0 \quad \dots(2.4.13)$$

From equation of continuity

$$\frac{AC}{V} = \text{constant}$$

Taking logarithm of both sides and differentiating

$$\frac{dA}{A} + \frac{dC}{C} - \frac{dV}{V} = 0$$

$$\frac{dA}{A} = -\frac{dC}{C} + \frac{dV}{V}$$

Substituting for  $\frac{dC}{C}$  from equation (2.4.13)

$$\frac{dA}{A} = \frac{VdP}{C^2} + \frac{dV}{V} = VdP \left[ \frac{1}{C^2} + \frac{1}{V^2} \left( \frac{dV}{dP} \right)_s \right]$$

Speed of sound in a fluid medium is given by

$$C_s = \sqrt{\frac{K}{\rho}} \quad \text{where } K \rightarrow \text{bulk modulus, } K = -V \left( \frac{\partial P}{\partial V} \right)_s$$

$$\rho = \frac{1}{V} = \text{density}$$

$$\therefore C_s^2 = VK = -V^2 \left( \frac{\partial P}{\partial V} \right)_s \Rightarrow \frac{1}{V^2} \left( \frac{\partial V}{\partial P} \right)_s = -\frac{1}{C_s^2}$$

Substituting in equation

$$\frac{dA}{A} = VdP \left( \frac{1}{C^2} - \frac{1}{C_s^2} \right)$$

For a finite process

$$\frac{\Delta A}{A} = V \Delta P \left( \frac{1}{C^2} - \frac{1}{C_s^2} \right) \quad (2.4.14)$$

In a nozzle we arrange a pressure drop down the duct continuously ( $\Delta P < 0$ ) velocity of the fluid should increase (i.e. an accelerated motion).

The question is what should the shape of the nozzle be. Consider the following cases in order :—

- 1) Initially as  $C < C_s$ ; velocity of the fluid is subsonic, r.h.s. of equation (2.4.14) is negative. So we require that  $\frac{\Delta A}{A}$  is -ve, i.e. the nozzle should be a convergent one (fig. 2.8).

- 2) As the velocity approaches the sonic velocity  $C = C_s$

$$\text{r.h.s.} = 0 \quad \therefore \Delta A = 0$$

The nozzle reaches a uniform cross section called its throat

- 3) Velocity increases further  $C > C_s$ ; r.h.s.  $> 0$ ,  $\frac{\Delta A}{A} > 0$ ;  $\Delta A$  should increase; the nozzle is a divergent one. Thus the proper geometry of the nozzle for a very high velocity steam jet is a convergent divergent nozzle. This is illustrated in fig. 2.8.

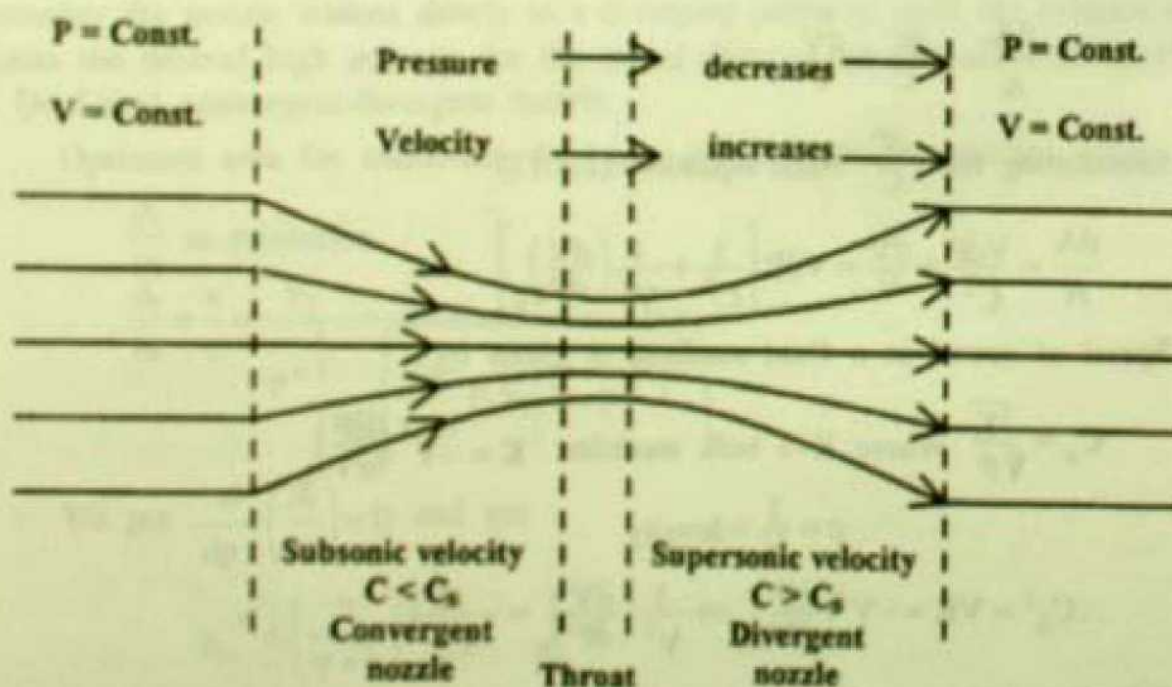


Fig 2.8 Throat

The same result could be arrived at from Mollier's diagram.



## Function of the turbine blades

The high velocity steam jet hits upon the turbine and suffers a change in direction. This produces an impulse and so a tangential force. The corresponding torque about the shaft rotates the turbine wheel.

By suitable speed governors a turbine may be used to produce power over a wide range of speeds. Efficiency also varies with the speed. To elaborate we refer to the graph (Fig. 2.9) that shows variation of torque and output power as functions of blade speed/input steam velocity.

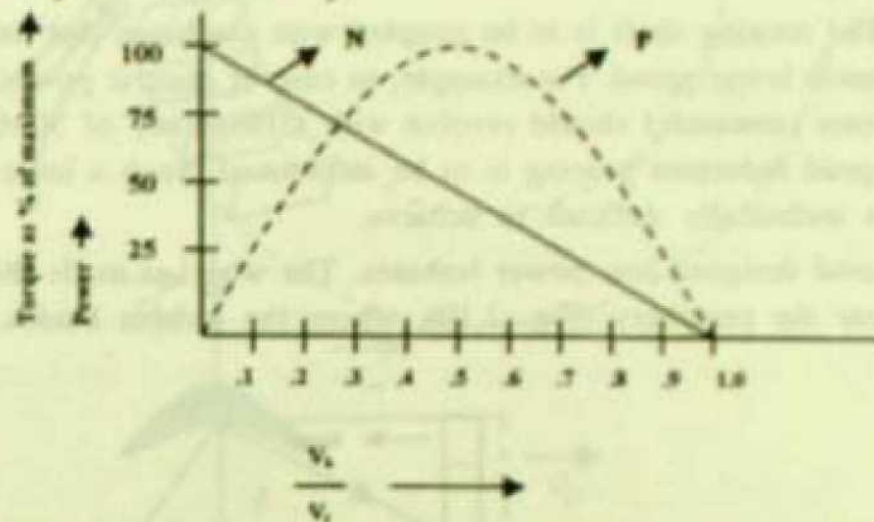


Fig. 2.9 Torque (N) and (P) of turbine blades

When the blade stalls, the jet enters and leaves with equal speed and produces maximum torque. However in that case output power,  $P=0$ . As the blades are allowed to speed up, the jet leaves more slowly and the torque drops. The power reaches maximum when the blade speed is half the speed of the incident steam and the steam leaves with a minimum speed. A simple calculation may help. We assume elastic collision

$C_i \rightarrow$  velocity of the incident steam jet

$C_f \rightarrow$  final velocity of the steam jet from the blading

$C_b \rightarrow$  tangential velocity of the blades

### For elastic collision

Relative velocity of approach = Relative velocity of separation

$$C_i - C_b = C_b - C_f$$

$$C_f = 2C_b - C_i$$

$$C_b = \frac{1}{2} C_i \quad \therefore C_f \rightarrow 0$$

The corresponding peripheral velocity of the turbine is around 1000 m/s, for typical velocity of incident steam jet used. This is very high speed and many practical difficulties arise, e.g.

- 1) For such speed, the angular velocity  $\approx 30,000$  rpm (considering radius of the turbine wheels in use). As such the centrifugal force becomes so large that the moving parts would fly into pieces.
- 2) Along with tangential force, axial force is also developed. This will cause mechanical vibrations.
- 3) The rotating shaft is to be coupled with machines that usually rotate with much lower speed. For example, in case of electric power generation, the rotor (armature) should revolve with a frequency of  $50 \times 60 = 3000$  rpm. So speed reduction gearing is to be introduced. Such a large speed reduction is technically difficult to achieve.

De Laval designed low power turbines. The wheel is made thick near the axis and thin near the periphery (Fig. 2.10), where the turbine blades are attached.

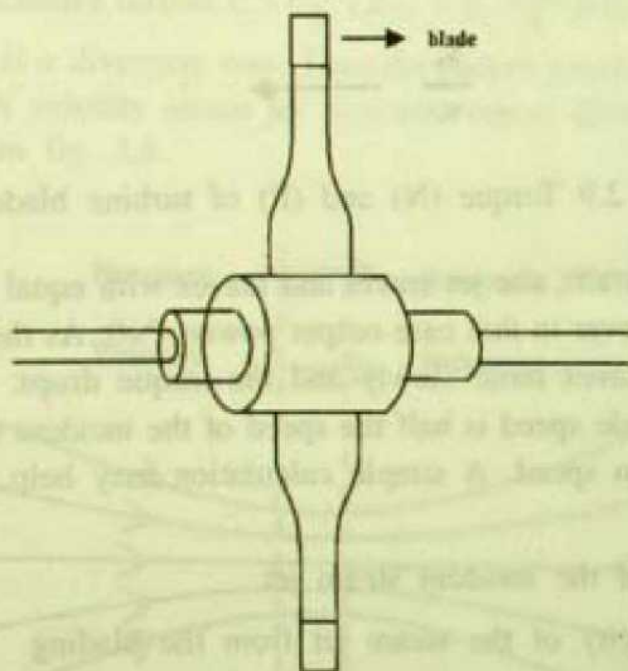


Fig 2.10 Front vertical section of turbine wheel (De Laval)

Figure 2.11 shows the velocity diagram of the incident steam jet from the nozzle, exhaust steam jet from the blade in relation to the turbine velocity ' $C_b$ ' at the periphery.

Fig. 2.11(a) shows the front view. The neighbouring turbine blades are mounted nearly parallel to each other. This offers little impedance to the steam jet. Steam pressure does not practically fall on its passage through the bladings.



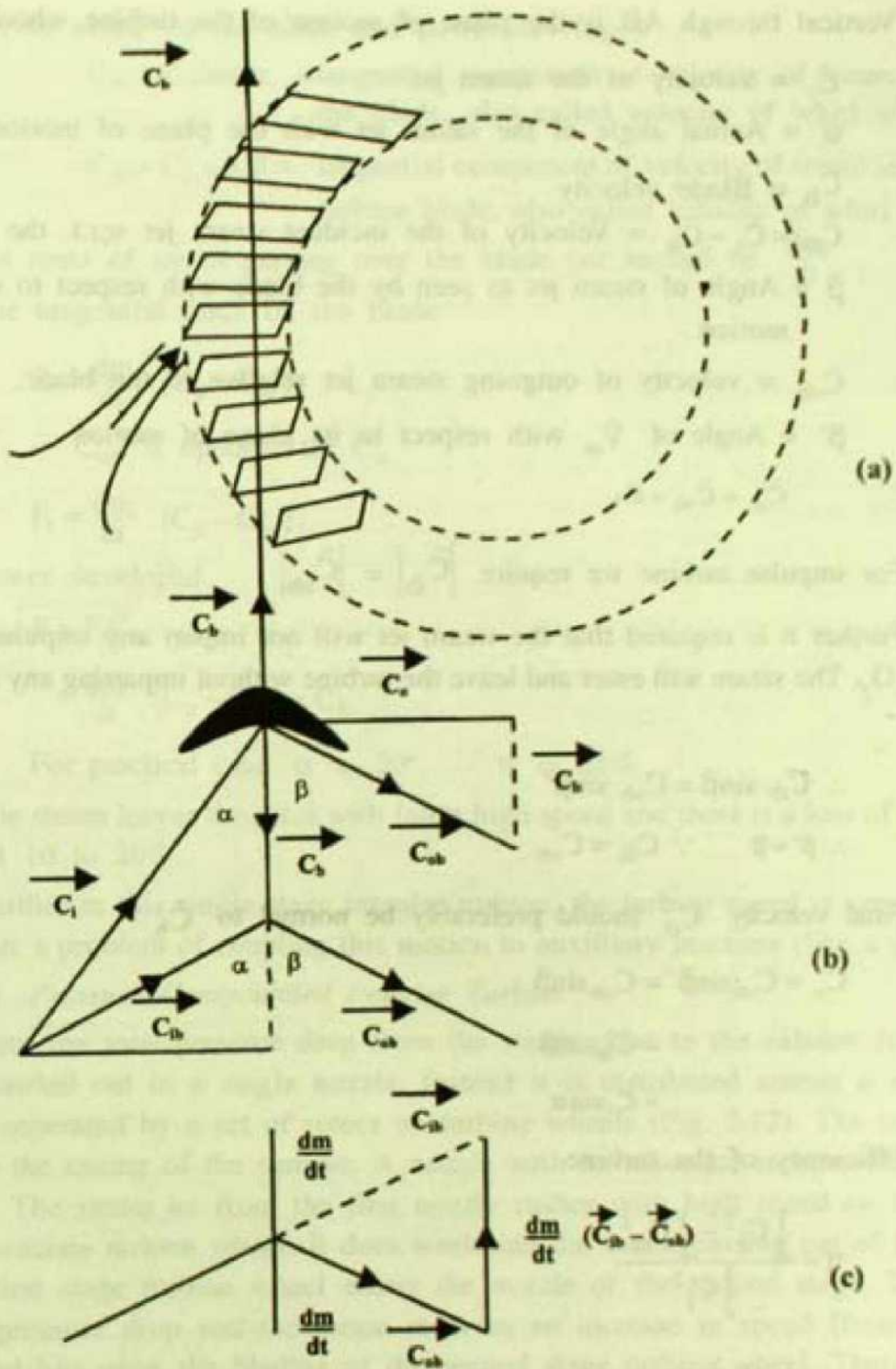


Fig 2.11 Velocity diagram

Fig. 2.11b is the velocity vector diagram and Fig. 2.11c indicates no axial component of force.

Vertical through AB is the plane of motion of the turbine wheel

$\bar{C}_i$  = velocity of the steam jet

$\alpha$  = Actual angle of the steam jet with the plane of motion.

$\bar{C}_b$  = Blade velocity

$\bar{C}_{ib} = \bar{C}_i - \bar{C}_b$  = Velocity of the incident steam jet w.r.t. the blade

$\beta$  = Angle of steam jet as seen by the blade with respect to its plane of motion.

$\bar{C}_{ob}$  = velocity of outgoing steam jet relative to the blade.

$\beta'$  = Angle of  $\bar{V}_{ob}$  with respect to its plane of motion

$$\bar{C}_o = \bar{C}_{ob} + \bar{C}_b$$

For impulse turbine we require  $|\bar{C}_{ib}| = |\bar{C}_{ob}|$

Further it is required that the steam jet will not impart any impulse along the axis  $O_1O_2$ . The steam will enter and leave the turbine without imparting any undesirable shock.

$$\therefore C_{ib} \sin \beta = C_{ob} \sin \beta'$$

$$\therefore \beta' = \beta \quad \because C_{ib} = C_{ob}$$

Final velocity  $\bar{C}_o$  should preferably be normal to  $\bar{C}_b$

$$\begin{aligned} C_o &= C_{ob} \sin \beta' = C_{ob} \sin \beta \\ &= C_{ib} \sin \beta \\ &= C_i \sin \alpha \end{aligned}$$

Efficiency of the turbine

$$\eta = \frac{\frac{1}{2} C_i^2 - \frac{1}{2} C_o^2}{\frac{1}{2} C_i^2}$$

$$= 1 - \frac{C_o^2}{C_i^2}$$

$$= 1 - \sin^2 \alpha$$

For maximum efficiency  $\alpha = 0$



Forces acting on the blade and power developed

$C_{it} = C_i \cos \alpha$  = tangential component of velocity of steam entering the blade, also called velocity of whirl at inlet.

$C_{ot} = C_o \cos \beta$  = tangential component of velocity of steam leaving the turbine blade, also called velocity of whirl at outlet.

Let mass of steam passing over the blade per second be  $\frac{dm}{dt}$

The tangential force on the blade

$$F_t = \frac{dm}{dt} (C_{it} - C_{ot})$$

$\therefore C_{ot}$  is opposite to  $C_{it}$

$$F_t = \frac{dm}{dt} (C_{it} - C_{ot})$$

Power developed

$$P = F_t C_b$$

$$= \frac{dm}{dt} (C_{it} - C_{ot}) \times C_b$$

For practical case  $\alpha = 20^\circ \therefore \eta = 88\%$

The steam leaves the rotor with fairly high speed and there is a loss of efficiency of about 10 to 20%.

Further in this single stage impulse turbine, the turbine speed is very high and this poses a problem of coupling this motion to auxilliary machine (like a generator).

### c) Pressure Compounded Impulse Turbine

Here the total pressure drop from the steam chest to the exhaust (condenser) is not carried out in a single nozzle. Instead it is distributed among a number of nozzles seperated by a set of rotors on turbine wheels (Fig. 2.12). The nozzles are fixed to the casing of the turbine. A nozzle with its associate turbine wheel forms a stage. The steam jet from the first nozzle rushes with high speed on the blades of its associate turbine wheel. It does work and the steam passing out of the blades of the first stage turbine wheel enters the nozzle of the second stage. There is a further pressure drop and the steam receives an increase in speed (from the heat drop) and hits upon the blading of the second stage turbine wheel. There are thus a number of stages. All the turbine wheels rotate about the same shaft. However, drop of pressure per stage is reduced. So velocity of steam leaving a nozzle and entering the associate blades is reduced. This helps in reducing the rpm of the wheel while still attaining higher power, distributed over a number of rotors. Energy loss is also substantially reduced. This is the principle of Reteaus turbine.

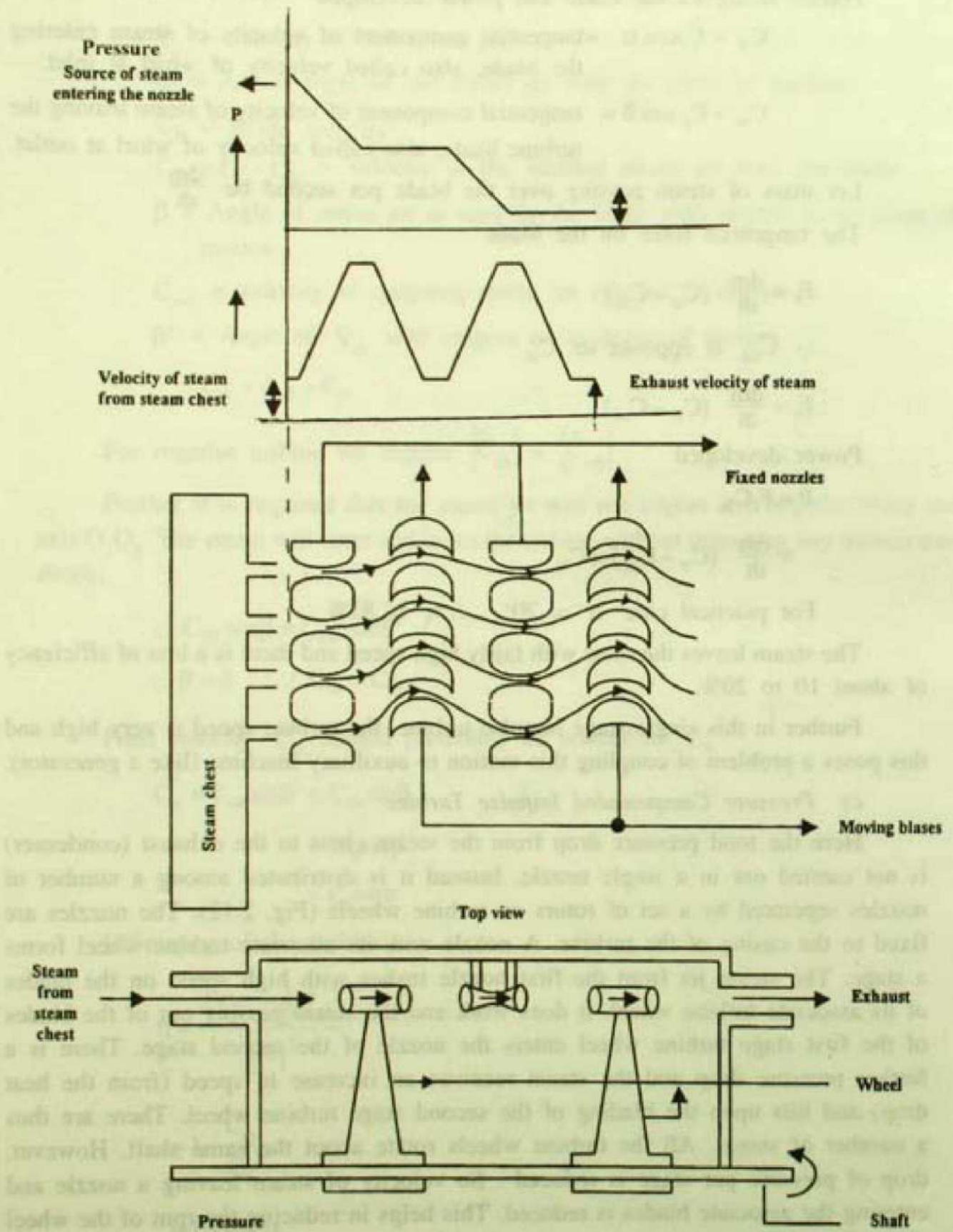


Fig 2.12 Pressure compounded impulse turbine



d) *Curtis turbine : Velocity compounded impulse turbine*

The principle of velocity compounding was designed by Curtis. It is essentially multi-impulse turbine. It consists of a set of nozzles (Fig. 2.13) followed by two or more sets of moving blades each set fixed on a wheel that can rotate about the common shaft. Between moving blades there are rows of fixed blades or guide blades fixed to the casing. The entire pressure drop is carried out in the nozzle (there are a set of these nozzles). The steam passes out through the nozzles and the steam jet attains a high ordered velocity. The steam jet from a nozzle through the ring of moving blades shares part of its kinetic energy with the first set of moving blades. The steam issuing out from the moving blades still has enough velocity and energy. It then passes through the fixed guide blades. These blades only change the direction of velocity of the steam and guide them to the next set of moving blades and a second share of kinetic energy with the wheels occurs. Thus the total kinetic energy of the steam is extracted by a row of moving blades in succession. Thus in each set the wheels gain only moderate energy and velocity and rotate with a high but moderate rpm. The wheels supply energy to the same common shaft. The shaft receives the compounded energy and yet rotates with a moderate rpm. The combined system possesses a very high moment of inertia and total energy ( $\frac{1}{2}I\omega^2$ ) received by the system is quite high. The angular frequency (rpm) produced is compatible with the rpm of the successive units that are to receive energy from the shaft.

**Velocity and pressure compounded turbine**

There is further modification where the principle of pressure compounding is also utilized. In this case the steam issuing out of the last stage of the moving blades possesses enough thermal energy. The steam is again made to pass through a second fixed set of nozzles, where they undergo further pressure drop and regain velocity and kinetic energy (from the heat drop). The energy is again extracted in steps by a second stage of array of moving and guide blades.

e) *Parson's Reaction Turbine*

It applies both the principle of reaction and impulse. Here there is no separate nozzle. The particular design of the moving blades makes them act like nozzles. There are alternate sets of fixed blades, fixed to the cylindrical casing of the turbine. In between them there are alternate sets of moving blades, fixed to the rotors (wheels) of the turbine. All the wheels rotate about the common shaft.

The fixed blades have a convergent passage and act as nozzles.

The moving blades also have a low convergent passage.

Steam from the steam chest enters normally and passes through the first set of (fixed) convergent blades. Here the steam undergoes some drop of pressure and its velocity increases. The steam then enters a row of moving blades. On passing through these blades two things happen (Fig. 2.14).



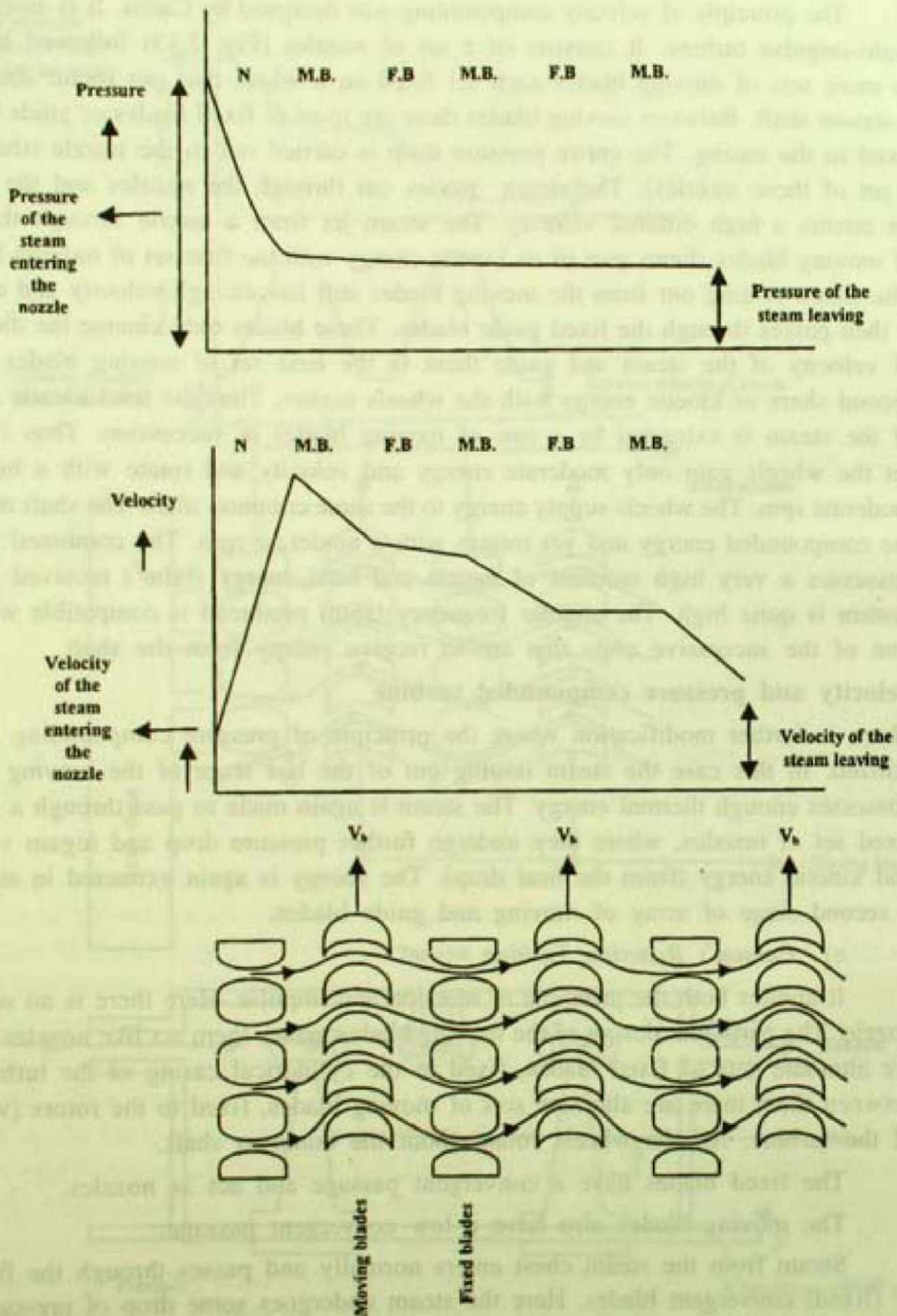
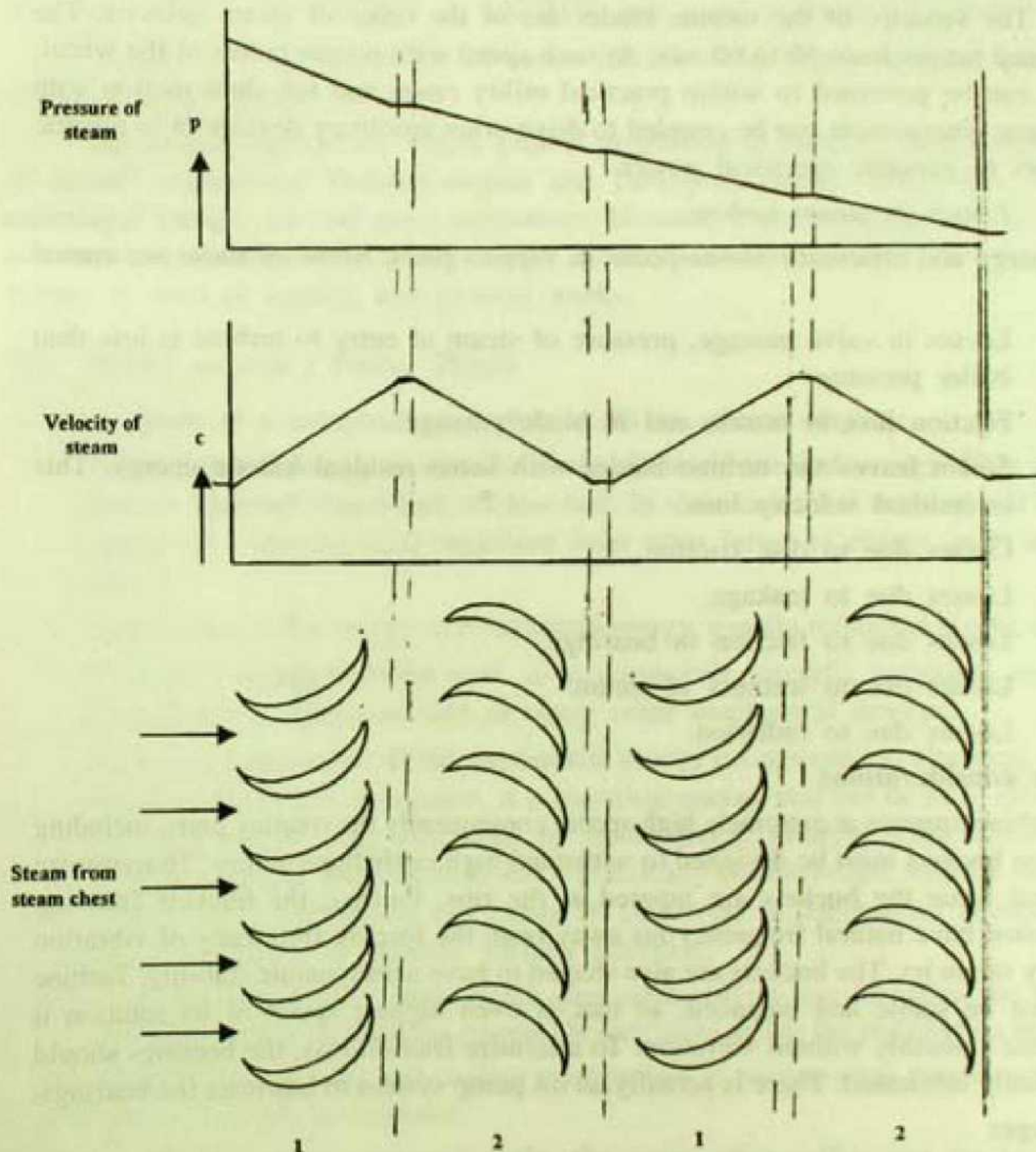


Fig 2.13 Velocity compounded impulse turbine



- 1) The direction of velocity of the steam changes, that produces an impulse on the wheel.
- 2) The moving blades being convergent causes a drop of pressure and thereby velocity and kinetic energy of the steam increase. This in, turn, produces a reaction on the wheel.



1 – fixed blades, 2 – moving blades

Fig 2.14 Reaction turbine

The impulsive force and the reaction force are made to act along the same direction, tangential to the wheel, in its plane of rotation and the wheel rotates.

In Parson's turbine there are usually a large number of stages of fixed and moving blades. Consequently total pressure drop is more or less equally distributed into a fairly large number of low pressure drops. Thus the steam velocity in each stage, entering the moving blades is significantly low, compared to that in De Laval's turbine. The velocity of the turbine blades are of the order of steam velocity. The blade speed ranges from 30 to 60 m/s. At such speed with proper radius of the wheel, the rpm can be governed to within practical utility range and the shaft motion with simple gear arrangement can be coupled to drive other auxilliary devices or to electric generators to generate electrical power.

f) *Losses in steam turbine*

Energy and efficiency losses occur in various parts. Some of these are named below.

- 1) Losses in valve passage, pressure of steam at entry to turbine is less than boiler pressure.
- 2) Friction loss in nozzle and in blade passage.
- 3) Steam leaves the turbine blades with some residual kinetic energy. This is residual velocity loss.
- 4) Losses due to disc friction.
- 5) Losses due to leakage.
- 6) Losses due to friction in bearings.
- 7) Losses due to wetness of steam.
- 8) Losses due to radiation.

**Machine considerations**

Steam turbine operates at extremely high speed, consequently the rotating parts, including the turbine buckets must be designed to withstand high centrifugal forces. To minimize centrifugal force the buckets are tapered at the tips. Further, the buckets (moving blades) must have natural frequency far away from the forcing frequency of vibration caused by steam jet. The buckets are also shaped to have aerodynamic stability. Turbine shaft must be stable and balanced, so that at even highest speed of its rotation it can operate smoothly without vibration. To minimize friction loss, the bearings should be constantly lubricated. There is actually an oil pump system to lubricate the bearings.

**Advantages**

Turbines have several advantages over the reciprocating engines, namely, (1) higher efficiency, (2) operation with less expensive fuel, (3) no upper limit to power output, (4) lower weight and small frontal area per unit output, (5) more stable balanced and less noisy operation, (6) no periodic change of temperature and negligible warm up time, (7) simple lubrication system (8) a long reliable life.



Turbines driven by steam are at present the most powerful and most widely used turbines. They can operate at constant and variable speed. High powered engines are almost always steam turbines. These are also ideal prime mover for driving machines that require rotation input power.

Most important application of steam turbine is generation of electric power. More than 85% of energy used these days is provided by steam turbine. A single unit with a rating of 500 MW to 1500MW and more is frequently used.

Second most important use of steam turbine is in ship propulsion (required power of about 75,000 H.P.)

Gas turbine engines are widely used in propulsion of aircrafts. Two major types of aircraft engines are Turbojet engine and Turbo-prop engine. Other uses include centrifugal pumps, air and gas compressors, blowers, paper machines etc. In many industries, steam turbines primarily supply motive power and exhaust from the steam turbine is used in heating and process works.

## 2.5 Power sources / Power Plants

A power plant is a composite unit that works in several stages :

- 1) Production of energy is usually in the form of thermal energy from energy sources like coal, fossil fuel, nuclear fuel. In some cases it is translational kinetic energy of a moving fluid organized from other forms of energy, as in a water wheel.
- 2) Conversion of this energy to mechanical energy, usually rotational kinetic energy.
- 3) Use of this energy to direct work, as in automobile, aircrafts, turbodrill, turbofan, mechanical compressors and in many other mechanical devices.

However, a major part of the mechanical energy is converted to electrical energy via generators. This is because, it is electrical energy that can be most efficiently transmitted over long distances by transmission systems with intermediate distribution grids. Such transmission is practically instantaneous. Also the electrical energy at the output end can easily be converted to mechanical energy (with large efficiency) to perform various activities.

- 4) A step up transformer with switch gear arrangement that increases the voltage (and lowers the transmission current) before being fed to the distribution network.
- 5) A suitable condensing and cooling arrangement in various steps mentioned above from (1) to (4) is required.
- 6) In case of a steam operated cycle, the steam after performing the necessary work, still contains thermal energy. Instead of recycling this steam, it is often used in different process plants like in paper mill, textile mill and in many others.



Mechanical / electrical power that is generated by a power plant must be distributed when it is generated. This is because there is no practical way of storing the power output and power cannot be traded for energy. This constraint results in a wide variation of loads imposed upon a plant. The required power must be made available when the load is imposed. Quite a large fraction of the capacity to produce power may be idle during long periods when there is no demand of power. This complicates the construction of the power plants and estimation of its efficiency. In view of this, apart from efficiency, we define two factors that are required to study the performance of a power plant.

$$\text{Capacity factor} = \frac{\text{average load for the period}}{\text{rated installed capacity}}$$

$$\text{Load factor}^* = \frac{\text{average load for the period}}{\text{Peak load in the period}}$$

Thermal efficiency of a power plant is limited by the maximum thermodynamic efficiency of the cycle. The standards of Carnot, Otto, Diesel and Rankine cycles are usual criteria to which efficiency is compared. However, these are ideal cycles, whereas actual operation involves various irreversible changes and heat losses. So the practical efficiency is appreciably smaller than the theoretical value. In this context it may be referred that the Rankine cycle is the simplest cycle, using condensable vapour, which is usually referred to judge the overall efficiency. It is used, because it offers a simple analysis even under the effects of varying steam generation pressure, temperature and condenser pressure.

The overall performance of a power plant from fuel to net usable output is expressed as thermal efficiency in % of fuel consumption (Lbwt or gallon per HP-hr or per kW-hr) or heat rate (BTu supplied in fuel per HP-hr or per kW-hr).

It may be noted that in case of a steam power plant, thermal efficiency can be increased both by raising boiler steam pressure and temperature. The average steam temperature in steam cycle is raised by either superheating or by raising the boiler pressure or both.

So, for economic output of a power plant, proper heat flow is to be maintained at various stages.

Therefore, a proper analysis requires a flow diagram or heat balance diagram. In figure 2.16 we sketch the basic flow diagram of different power plants.

We now describe in brief two power plants with turbine as the prime mover.

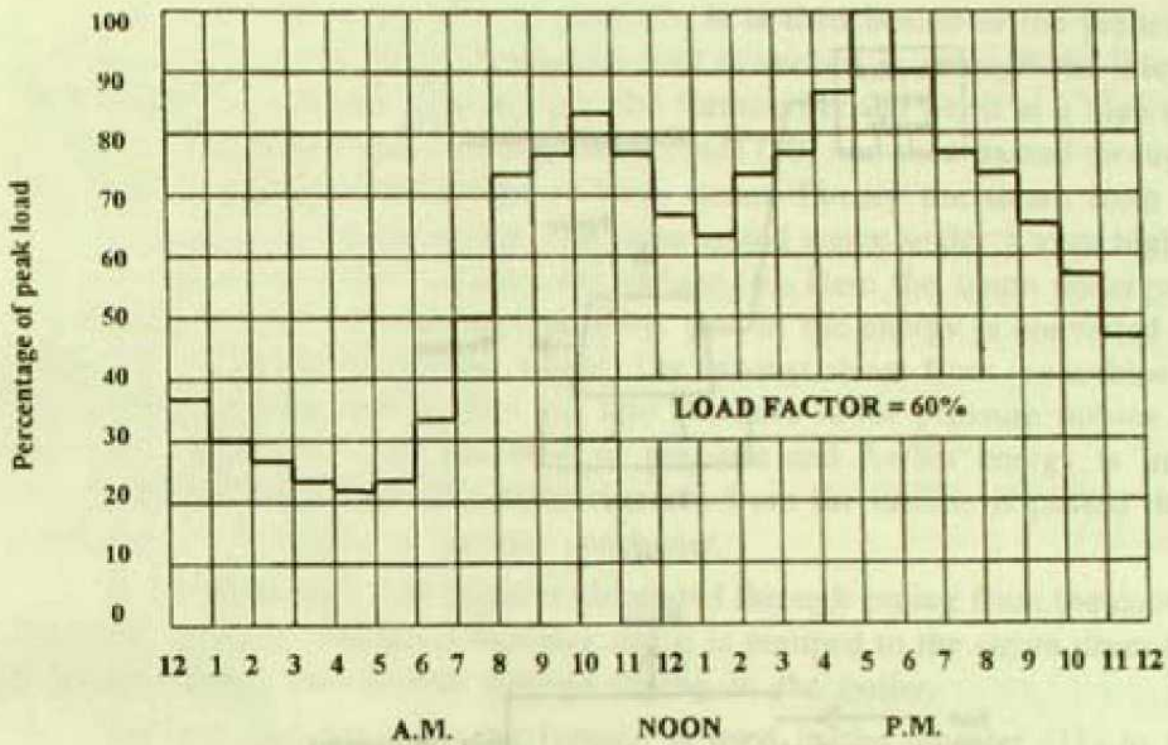
#### a) *Steam power plant*

Figure 2.17 shows a coal-fired, steam turbine operated, electricity generating power plant.

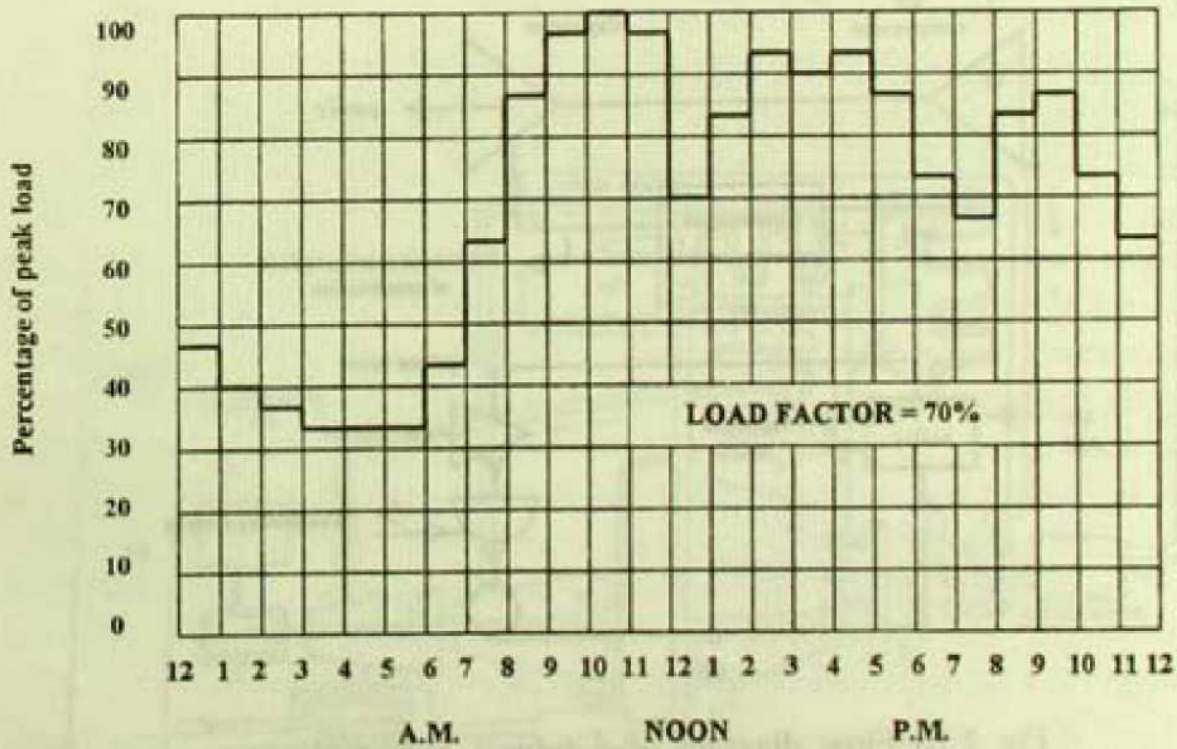
---

\* Fig 2.15





TYPICAL DECEMBER UNIT LOAD CURVE OF A METROPOLITAN SYSTEM LOAD



TYPICAL AUGUST UNIT LOAD CURVE OF A METROPOLITAN SYSTEM LOAD

Fig 2.15

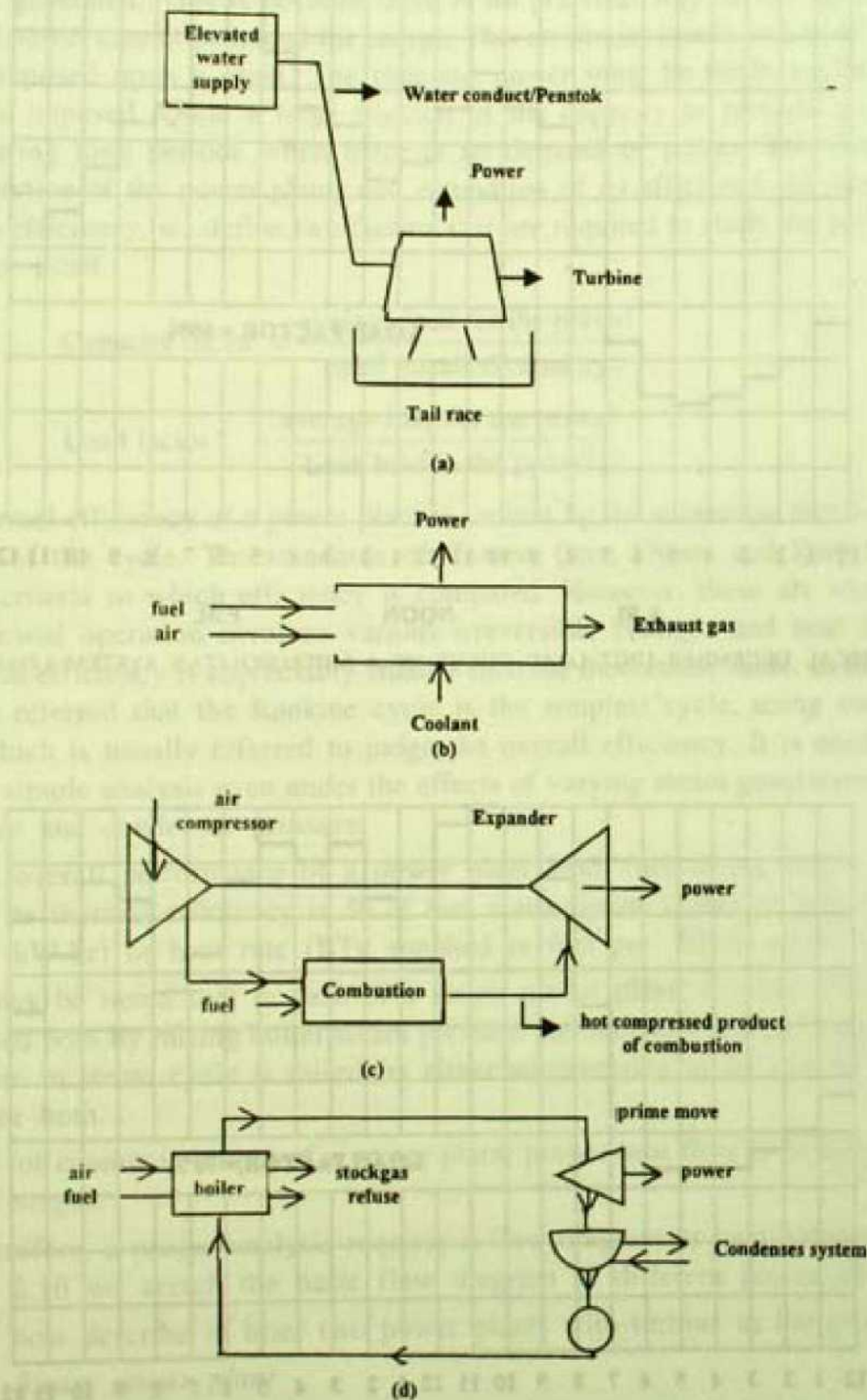


Fig 2.16 Flow diagram of a typical power plant  
 a) Hydel power plant, (b) I-C engine power plant,  
 (c) gas-turbine, (d) fossil fuel powered power plant



Air is drawn in through the pipe (1). It is then heated in the preheater (2) by the flue gas from the boiler. Pulverised coal is sucked in through the inlet (3). Hot air mixed with coal dust is blown into the furnace (4) and burnt at a high rate under the boiler (5). Water is taken in the steam drum (10) and then passed through several kilometer of piping in the boiler to form steam. Finally the steam from the drum is heated at the top of the boiler. The superheated steam under a very high pressure and temperature is passed through the turbine (6). Here the steam undergoes a drop of pressure and due to resulting heat drop, part of the energy is converted to kinetic energy of rotation of the turbine wheel. The exhaust steam from the turbine is passed through the reheater and is then fed into the next lower pressure turbine (7). Here the steam undergoes a second drop of pressure and further energy is imparted to the second turbine wheel. The exhaust steam from the turbine is passed through the condenser. It is usually a surface condenser.

In the condenser cold water is circulated through piping from the cooling tower (14). The steam is condensed to water and it is pumped to the steam drum (9), where it is once again recirculated through piping in the boiler.

The hot flue gas from the furnace is used in the reheater (11) to reheat the exhaust steam from the turbine (6). The flue gas is also used to heat the air in the chamber (2). The hot gas is then passed through the electrostatic precipitator (15) and is finally allowed to escape through the chimney (16).

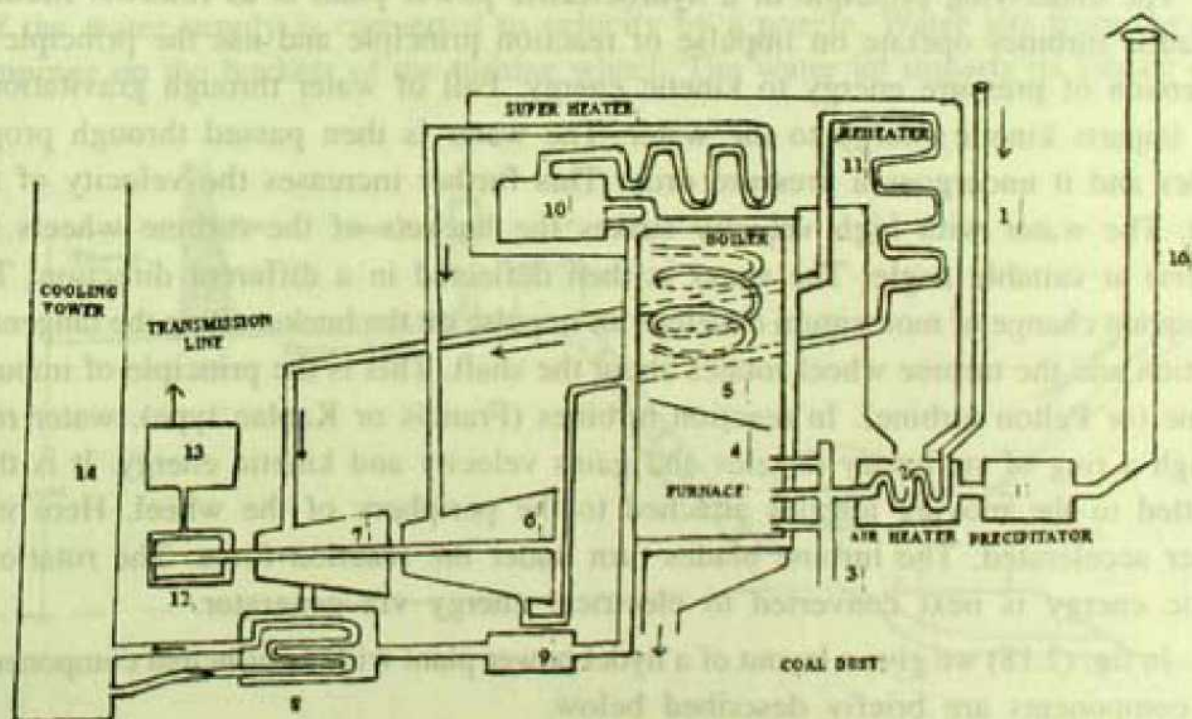


Fig 2.17 Steam turbine operated power plant.



The turbine wheels in (6) and (7) rotate about the common shaft. The shaft drives the A.C.generator (12) at about 240Volt, 50Hz. The output is then stepped up to about 500kV by the step up transformer (13) and power is finally distributed by the transmission line.

#### a) *Hydroelectric Power Plant*

Hydel power is known since early nineteenth century and with our knowledge of conversion of mechanical energy to electrical energy, hydel power stations are being widely installed.

The first hydroelective power station in India had been established in 1897, at Sidrapong in Darjeeling. It started with a plan capacity of only 130 kW which at present is 600 kW. The present installed capacity of hydropower in West Bengal is given as follows

Jaldhaka Hydel Project	— 35 MW
Rammam Hydel Project	— 51 MW
Teesta Canal fall hydropower	— 67.5 MW
Fazi power station	— 2 MW
Sidrapong	— 0.6 MW
Little Rangit Project	— 2 MW
<hr/>	
158.1 MW	

The underlying principle of a hydroelectric power plant is as follows. Modern Hydraulic turbines operate on impulse or reaction principle and use the principle of conversion of pressure energy to kinetic energy. Fall of water through gravitational head imparts kinetic energy to the water. The water is then passed through proper nozzles and it undergoes a pressure drop. This further increases the velocity of the water. The water with high velocity strikes the buckets of the turbine wheels (or runners) at suitable angle. The water is then deflected in a different direction. The consequent change of momentum develops an impulse on the bucket along the tangential direction and the turbine wheel rotates about the shaft. This is the principle of impulse turbine (or Pelton turbine). In reaction turbines (Francis or Kaplan type), water runs through a ring of stationary nozzles and gains velocity and kinetic energy. It is then admitted to the moving nozzles attached to the periphery of the wheel. Here it is further accelerated. The turbine blades turn under the reaction force. The rotational kinetic energy is next converted to electrical energy via generator.

In fig. (2.18) we give a layout of a hydel power plant with its principal components. The components are briefly described below.

#### 1) *The Dam*

A dam is constructed across the path of the river. For economy we need to



keep the length of the dam short. So the site is selected near the neck of the river. The river valley on its upstream side provides the large storage system. If possible the site is selected a little beyond the confluence of two rivers. In such a case both the river valleys contribute to a large storage capacity. The two principal functions of a dam are

- (i) to build up a large storage area and
- (ii) to provide a suitable water head required for the turbine.

## 2) *Spill ways*

It releases the excess water due to heavy rainfall or flood or both and helps to maintain waterhead below a predetermined upper limit.

## 3) *Pressure channel*

The water from the reservoir runs through a pressure channel and enters the penstock (discussed below) through an inlet gate, where screens are placed to prevent unwanted objects entering the turbine.

## 4) *Surge tank*

It is used to regulate the pressure and velocity of water in the penstock.

## 5) *Penstock*

It is a pipe moving downward a little beyond the surge tank to the turbine. It is made up of steel tube through reinforced concrete.

## 6) *Turbine*

Water from the bottom of the penstock is fed into the turbine. The turbine is either impulse or reaction type. In the impulse turbine (Pelton turbine) pressure of the water supply is converted to velocity by a nozzle. Water jets from the nozzle impinge on the buckets of the turbine wheel. The water jet imparts its kinetic energy

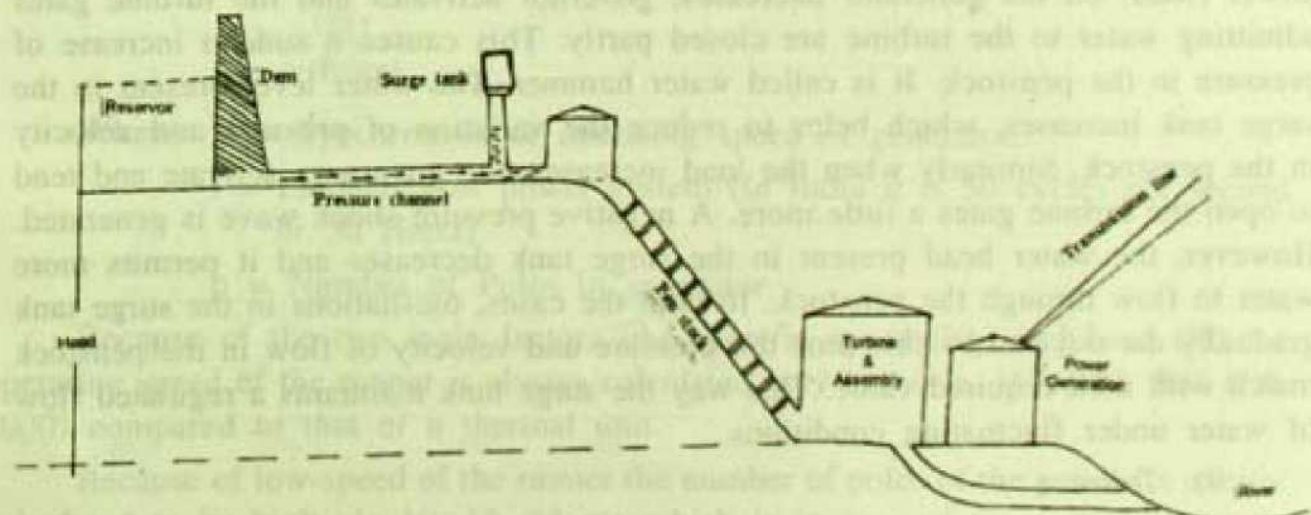


Fig 2.18 Layout of a hydel power plant.



to the buckets. Impulse turbines are usually operated from high head source. Such turbines with moderate head use horizontal shaft. However, for large head installations, about 4 to 6 jets operate on a single wheel. In such cases vertical shaft is favoured because it involves lesser friction on bearings. Reaction type turbines are Francis turbines or Kaplan turbines. In a reaction turbine the chamber of a runner (rotating part of the turbine) is completely filled with water coming from the Penstock. The reactive pressure of the water acts upon the runner and rotates the runner thereby releasing the water pressure. The released water from the runner passes out through the draft tube and falls in the tail-race. The shaft of such turbine should be vertical. Water heads for different turbines are given below :

Pelton turbine	high heads around 300 m
Francis turbine	medium heads around 60 m
Propeller or Kaplan turbine	low heads around 15 m.

#### 7) *Power house*

All the turbine wheels rotate about a common shaft. This rotational kinetic energy is used to drive an alternator and electrically energy is produced. The generators are usually 2-pole and the frequency of supply is 50 Hz, with fluctuations within  $\pm 0.05$  Hz. Thus the generator rotation speed should be maintained at 3000 rpm. The shaft rotation speed is higher. So proper reducing gear arrangement is required for coupling the shaft to the alternator.

Output power from the generator is fed to a step up transformer and electrical power at a very high voltage and low current is fed to the grid of the power transmission line.

#### 8) *Governors and function of surge tank*

To maintain a stable and synchronous speed, the turbine motion is always coupled to a speed governor as an integral part of its operation. When demand of output power (load) on the generator decreases, governor activates and the turbine gates admitting water to the turbine are closed partly. This causes a sudden increase of pressure in the penstock. It is called water hammer. The water level present in the surge tank increases, which helps to reduce the variation of pressure and velocity in the penstock. Similarly when the load increases, the governors activate and tend to open the turbine gates a little more. A negative pressure shock wave is generated. However, the water head present in the surge tank decreases and it permits more water to flow through the penstock. In both the cases, oscillations in the surge tank gradually die out and by this time the pressure and velocity of flow in the penstock match with their required value. This way the surge tank maintains a regulated flow of water under fluctuating conditions.

#### 9) *Tailrace*

The water from the turbine enters the tail race channel and finally joins the river on its downstream course.



## NORMAL SPEED OF THE RUNNER

Unlike speed of the thermal-unit (which most often is 3000 rpm), speed of the hydro-unit is related to its water-head and the power it is designed to generate.

The normal or operating-speed ( $N$ ) of a turbine-runner may be determined from the preliminary selection of a suitable specific speed.

Specific-speed is defined as the speed of a symmetrical or homologous runner when it has been so reduced in size that it develops 1kW power under 1 meter head.

$$\text{Mathematically, } N_s = \frac{N\sqrt{P}}{H^{5/4}}$$

Where  $N_s$  = Specific Speed

$N$  = Operating Speed

$P$  = Power

$H$  = Head

Specific head is same for all capacities of turbine operating at same head. (Fig. 2.19).

To find the speed of the runner we can rewrite the equation as follows

$$N = \frac{N_s H^{5/4}}{\sqrt{P}}$$

The value of normal operating speed of runner  $N$ , thus obtained from above equation, shall be corrected suitably to match the synchronous speed of generator which is given by the formula

$$N = \frac{120 f}{p \text{ (Poles)}}$$

Where  $N$  = Synchronous or operating speed of generator.

$f$  = Frequency of power system (In India it is 50 cycles per second or 50 Hertz)

$p$  = Number of Poles in generator

Because of the two main factors like specific speed ( $N_s$ ) and head ( $H$ ) the operating speed of the runner is always calculated very low (viz. 167, 250, 500, 600, 1000) compared to that of a thermal unit.

Because of low-speed of the runner the number of poles of the generator exciter calculated to be high viz 10, 12, 24 etc which is in case of thermal exciter only two. For this reason poles of the exciter of hydro-generating unit are to be made salient ones (unlike cylindrical exciter in case of thermal unit).

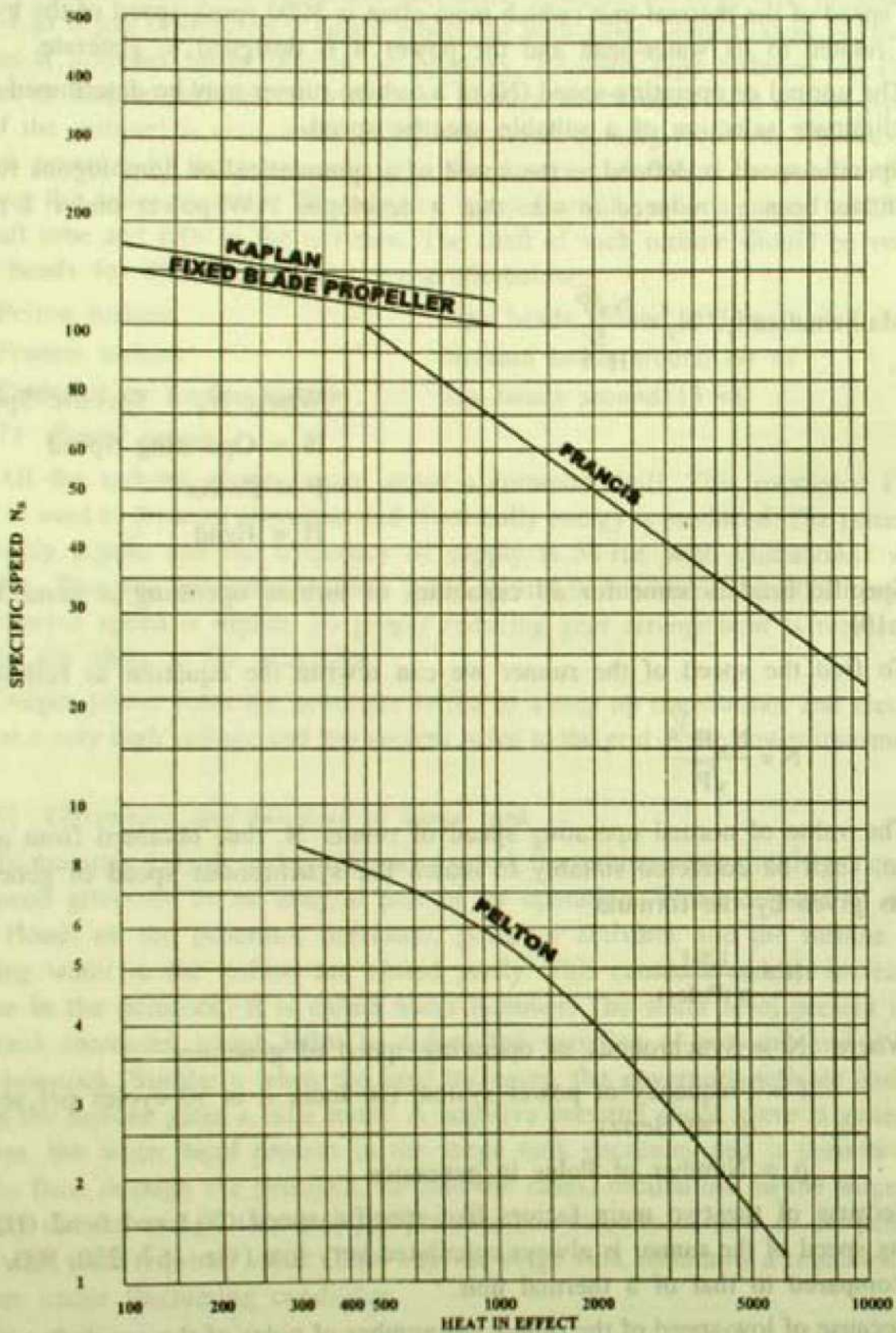


Fig. 2.19 Recommended upper limits of specific speeds for turbines for various rated speeds



## Conclusion :

Almost all hydro-power stations are situated in the remote place of hilly area beyond the load centre. Its infrastructural cost is more than that in case of thermal power station. The proportion of construction cost for civil and that for electro-mechanical is 60 : 40.

Hydro-power is being given much importance because of its low cost and pollution-free environment.

In our country installation of Micro-hydel and Mini-hydel power projects are also being envisaged.

The capacity below 25MW upto 1MW is called Mini Hydel whereas capacity below 1 MW is called Micro-hydel.

Two special types of plants are described below.

### *Diversion Canal Plant*

Such types of plants are in use where the river valley has a steep slope. The waters of the river are diverted away through a by-pass canal or power canal. A power plant is situated at a suitable location on this canal. The water after passing through the power plant joins the parent river. In North Bengal the ground has a typical slope. The Teesta Canal Fall Hydel Project is aimed to construct three power stations each having an installed capacity of 22.5 MW. Each power station has three penstocks each with installed capacity of 7.5 MW, 22.3km and 31.5km along the Mahananda main canal. The amount of fall are 7M, 4.65M and 4.64M respectively.

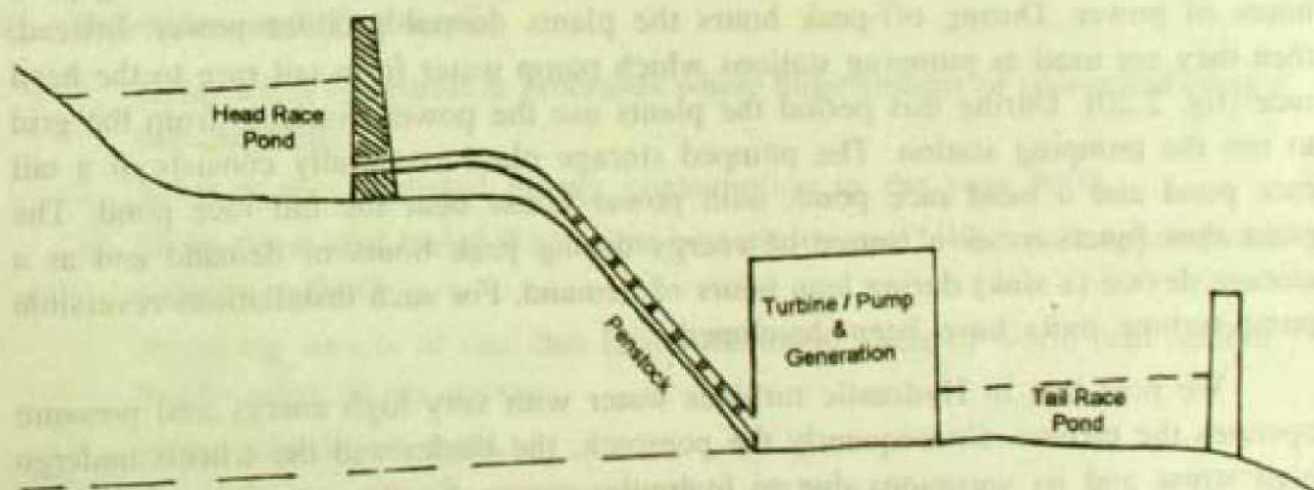


Fig 2.20 Pumped storage hydel plant



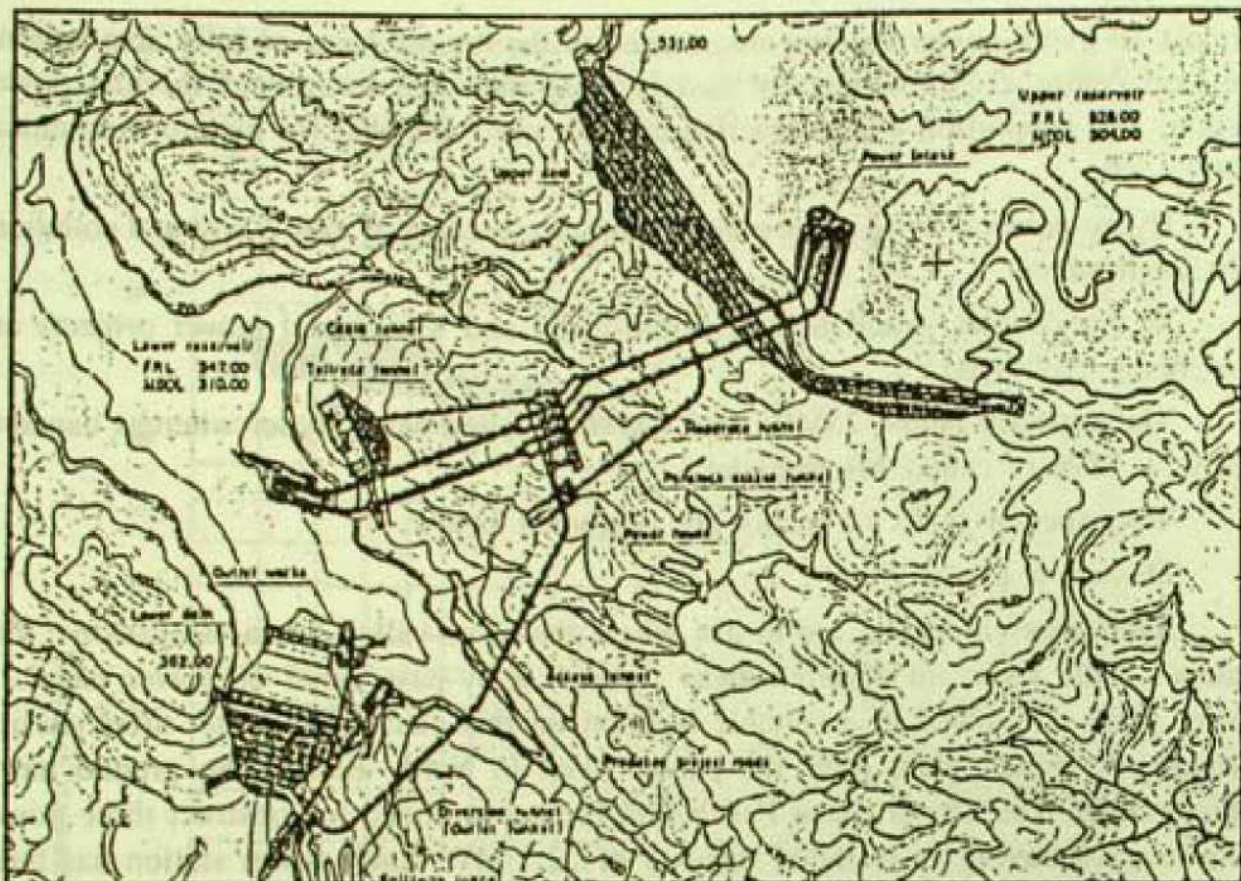


Fig. 2.21 Purulia Pumped Storage Project

### *Pumped Storage Plant*

Such plants are ordinary hydel power plants for part of the time during peak hours of power. During off-peak hours the plants do not produce power. Instead, then they are used as pumping stations which pump water from tail race to the head race (fig. 2.20). During this period the plants use the power available from the grid to run the pumping station. The pumped storage plant essentially consists of a tail race pond and a head race pond, with power house near the tail race pond. The plant thus functions as a source of energy during peak hours of demand and as a storage device (a sink) during lean hours of demand. For such installations reversible pump-turbine units have been developed.

We note that in Hydraulic turbines water with very high energy and pressure operates the turbine. Consequently the penstock, the blades and the wheels undergo high stress and its variations due to hydraulic surges. So the material, particularly for the bucket and the turbine wheels should be selected and designed to withstand such high stress and strain. The material should also be resistant to erosion and corrosion.

Owing to variation of load the shaft tends to vibrate. The speed at which such load variations produce transverse vibrations whose frequency match the natural frequency



of transverse vibration or some higher harmonic are called critical speed. At such speed transverse vibrations will produce resonant oscillation which is highly undesirable. The shaft may break or buckle. Hence it is necessary that the shaft should be carefully balanced and actual speed of operation should be well away from critical speeds.

All these technical difficulties are properly met and in modern Hydraulic installations the efficiency is very high, more than 85%.

To summarize, a hydel project involves huge civil, mechanical and electrical installations with advanced technology. It also requires a vast area of land. Before the construction rigorous survey of the land and its habitation is necessary. Factors like slope of the valley, catchment area, amount of rainfall, its yearly and monthly variation, downstream flow pattern should be carefully considered. The land and its surroundings must not be prone to earthquakes. The construction and design of the dam, and other components must be stable and fault-free as far as practicable, so as to be capable of absorbing the sockwaves resulting from sudden surge of water due to heavy rainfall and flood. Such construction also involves huge area of land and suitable rehabilitation of displaced people. All these add up to a huge expenditure and in short time operation cost per kW appears to be very high. However, in such projects the source of energy, the water power is inexhaustible and is being continuously replenished by the Sun and our weather cycle. Also in a developed hydel project, power production is a part activity only. It develops into a multipurpose project. It helps to distribute our water reserve more or less uniformly over the year and to bring out a planned irrigation scheme over a wide adjoining area. Such projects also help in flood control. Other activities like pisci culture, tourism can be developed. These are ecofriendly and can help the people over a wide region surrounding it.

## 2.6 Questions and Problems

1. Name a few industries & processes where huge amount of power and energy are required.
2. What is the estimated energy consumption in the year 2000.
3. Name three solid fuels? What is the best quality coal? What is its approximate calorific value?
4. Name the variety of coal that forms the major share of world coal reserve. Name some of its uses.
5. Name two liquid fuels.
6. Name some of the important families of hydrocarbons.
7. What are the basic processes from which different fuels can be obtained from crude oil.
8. Name a major known source of liquid hydrocarbons.
9. Name the major derivative from refined oil and mention their uses.



10. Name four commonly used gaseous fuels.
11. Define Calorific value of a fuel.
12. What are higher calorific value and lower calorific value of a fuel? How are they related?
13. Give two broad classification of nuclear fuels.
14. How energy is released in (i) Fission Process & (ii) Fusion Process?
15. Name the principle devices of Energy Storage
16. What characteristic properties should a material have for sensible heat storage?
17. Define volumetric heat capacity.
18. How flywheels are used to store energy?
19. What is pumped hydraulic storage?
20. What is the principle difference between the operations of an IC engine & turbine?
21. Give the principle of operation of a) Impulse Turbine & b) Reaction Turbine
22. Consider a thermodynamic system? Consider external kinetic energy, external potential Energy & Work done by the system on a machine, called the shaft work. For such a system write the 1st Law of thermodynamics.
23. Apply the above equation (Prob. 22) to study the steady State flow of a fluid through a device when the fluid enters a system through a section at a pressure  $p_1$ , with a velocity  $C$ , and at an elevation  $Z_1$ . The fluid leaves the system through another section where the corresponding parameters are  $p_2$ ,  $C_2$  &  $Z_2$  respectively. Consider unit mass of the fluid. Take  $U_1$ ,  $V_2$  &  $U_2$ ,  $V_2$  as internal energy & volume respectively at the entrance & exit section.

Set up an expression for the work done per unit mass of fluid considering the shaft work  $W_s$ . Also establish that, the heat absorbed

$$Q = U_2 - U_1 + \frac{1}{2}(C_2^2 - C_1^2) + g(Z_2 - Z_1) + (P_2 V_2 - P_1 V_1 + W_s)$$

24. What is a nozzle? A fluid under high pressure passes through a nozzle, where a pressure drop is maintained by some external device. Show that heat drop is equal to the gain in kinetic energy of the fluid. State the assumptions you make.
25. Assume isentropic flow of fluid (assumed to be perfect gas) through a nozzle

and show that the heat drop =  $\int_{P_1}^{P_2} V dP$ , where  $P_1$  &  $P_2$  are the pressures at the inlet and the exit end.



26. Use the above equation of problem 25 to establish that

$$C_2^2 - C_1^2 = 2(h_1 - h_2) = \frac{2\gamma}{\gamma-1} P_1 V_1 (1 - P^*)^{\frac{\gamma-1}{\gamma}}$$

where  $P_2/P_1 = P^*$ . Other symbols have usual meaning.

27. Use the equation in previous problem and explain how for a given  $P^*$  you can calculate
- Specific volume of steam at any section
  - Required area of cross section of the nozzle and
  - Speed of the steam
28. Establish De Laval's theory of convergent-divergent nozzle, so that the steam issuing out from the nozzle attains a high ordered velocity for a pre-set drop of pressure.
29. Assume equation 2.42 and also assume that no shaft work is done at any section. Now considering the equation of continuity establish that

$$\frac{\Delta A}{A} = v \Delta P \left( \frac{1}{C^2} - \frac{1}{C_s^2} \right)$$

&  $C_s$  is the sonic speed.

Hence set up the theory of convergent-divergent nozzle so that the steam issues out with a high ordered velocity.

30. Draw a sectional diagram for the geometry of the convergent-divergent nozzle for a very high velocity of the steam jet.
31. Steam from the nozzle hit upon the turbine blades and suffers change in momenta. Draw a graph showing the variation of torque and power developed by the steam jet as a function of  $C_b/C_i$ . Where  $C_b$  = Tangential velocity of the blade &  $C_i$  is the incident velocity of steam jet.
32. Assume elastic collision of steam jet with a blading and show that, steam leaves with minimum speed and the power developed is maximum when  $C_b = \frac{1}{2} C_i$ .
33. Draw suitable section of the blades and draw the velocity vector diagram of the steam jet. Calculate the efficiency of the turbine.
34. Set up an expression for the power developed in terms of the tangential component of velocity of the steam, entering & leaving the turbine & the blade velocity.
35. Explain the principle of operations of a pressure compounded impulse turbine. Draw graphs showing the variation of pressure & velocity of steam as it passes through the fixed nozzle and the moving blades.

36. Explain (i) Operations of velocity compounded impulse turbine and  
(ii) Parson's Reaction turbine
37. Power plant is a composite unit that works in several stages. Name the stages.
38. Can we trade Kilowatt hour for Kilowatt — explain.
39. What is the form of energy that can be most efficiently transmitted over long distance?
40. Define capacity factor load and factor of a power plant.
41. Name the standard cycle used to compare the overall efficiency of a power plant.
42. What is thermal efficiency regarding the performance of power plant?
43. Draw the heat flow diagrams of  
a) Hydel power plant, b) I. C. engine power plant, c) Gas turbine power plant, d) Fossil fuel fired power plant.
44. State the principle of operation of Hydel power plant.
45. Give a layout of Hydel Power Plant & label the principal parts.
46. Describe in brief, i) Diversion Canal Plant, (ii) Pump Storage Plant.

## 2.7 References

1. Energy (vol : I, II & III) : S. S. Penner and L. Icerman. Addition - Wesley Publishing Company Inc.
2. A. Treatise on heat : M. N. Saha and B. N. Srivastava, The Indian Press (Publication) Private Ltd., Allahabad.
3. Thermodynamics, Kinetic theory and Statistical Thermodynamics : Francis W. Sears & Gerhard L. Salinger, Narosa Publishing House, New Delhi.
4. Fundamentals of Classical Thermodynamics : J. Gordon and Richard E. Sontag, Wiley Eastern Ltd.
5. Thermal Engineering : P. L. Ballaney, Khanna Publisher
6. Mc Graw-Hill Encyclopaedia of science and technology.
7. West Bengal State Electricity Board Annual Review (1999).
8. Sechpatra (August 1998); Irrigation and Waterways Department Government of West Bengal.
9. Generation, Distribution and Utilization of electricity G. L. Wadhwa; Revised edition; Wiley Eastern Limited.
10. Private Communications, J. K. Saha (WBSEB).



## Chapter 3

### Non-Conventional Energy Sources

#### 3.1 Types of energy

Energy is required, by definition, to do any kind of work; the rate at which it is used is measured as power. We can do a certain amount of work slowly, using little power or quickly using more power involving the use of a similar total amount of energy in either case.

Virtually all human activity requires work to be done:— ploughing fields, lifting water, transporting ourselves, mining for minerals, manufacturing, cooking food and so on. Even lazing around and doing nothing involves “work” in the technical sense, as our bodies consume a minimum amount of energy to keep our heart, lungs and other vital organs functioning. Our own muscles, or the muscles of domesticated animals were for centuries the primary source of power behind human industry, and this remains so till this day for the majority of the human race. However, men or women power alone is hopelessly unproductive as a prime mover, for example, one man-year of hard labour represents a mere 150 kWh, which is the energy in about 15 litres of kerosene. You can buy 15 litres of kerosene at Rs. 50/- but one man year costs more than Rs. 10,000.00 assuming Rs. 30.00 as his daily wage. The first mechanical energy was produced by wind or water but these machines disappeared with invention of steam engine. Coal and oil became prime source of energy. Coal was the main source of energy upto 1958. The oil took over the control of energy domain in the last 50 years and oil is now the main source of energy. Today 6 billion tons of oil are used every year. The known petroleum reserves of our earth are estimated at 90 billion tons. As the supply of oil decreases the price of oil could be so high that the use of this source of energy by common people will be virtually impossible.

#### Coal

Reserves of coal are still fairly large, mainly in Russia, the USA and China. Experts estimate that these amount to 7600 billion tons. The coal will be utilised upto 2050 to reach its maximum use about the year 2150 and will not become scarce until the year 2300. The extraction of coal is getting more difficult; fifty years ago the depth was 350m, today the average depth is from 800m to 1000m.

#### Atomic Energy

This is one of the greatest hopes of this century. Already in America 10% and in Europe 5% of the energy requirements are met from atomic sources. Some scientists predicted



that in the year 2010 about 70% of the world energy requirement will be met from Atomic Energy. This has, however, proved to be unrealistic. However, the unsolved technical problems are very great and theoretical results cannot be put into practice.

The world is becoming increasingly conscious of the way that it uses energy and of the ways in which it might sustain development in the future. Renewable energy is a vital consideration. The price of electricity from renewable energy has fallen dramatically worldwide. Some technologies are competitive in the open market. There is no doubt next century will be the century of Renewable Energy Sources. It is estimated that in the middle of the next millennium 75% of the world energy required will come from Renewable Energy Sources. This is not only due to the fact that Renewable Source of Energy is inexhaustible but also for its contribution towards *protection of environment*. The threat posed to *sustainable* energy development by the increase in greenhouse gas emission and the consequent climate change occurring globally has understandably created worldwide concern. As concern for the environment grows, the prospects for renewable energy sources improve worldwide. The main Renewable Energy Sources are:—Solar Energy, Wind Energy, Energy from Biomass Resources, Energy from Falling Water, Energy from Tides, Geo thermal Energy etc.

### **The Sun and Solar Energy**

The origin of most of the energy available on the earth is the Sun. The interior of the Sun is inaccessible for experiments. However, based on observations of the solar surface and theoretical considerations, the interior temperature of the Sun is considered as 15 million Kelvins. In the interior of the Sun, enormous amount of energy is generated continuously due to nuclear fusion of 4 million ton hydrogen per second to helium. This energy comes to the surface of the Sun i.e. photosphere and is eventually emitted into space in the form of electromagnetic radiation.

The earth only receives a fraction of the energy generated at the interior of the Sun. This energy is of the order of  $1.8 \times 10^{11}$  MW and around 27000 times the energy produced by all human made systems in the world. Above the earth's atmosphere, the energy in sunlight is  $\approx 1358 \text{ W / M}^2$ . At sea level, this energy is reduced by atmospheric attenuation [absorption by water vapor,  $\text{CO}_2$ , dust particles etc.] to  $\approx 930 \text{ W / M}^2$ .

### **3.2 Solar Energy**

Solar Energy can be harnessed in two different routes:

(a) Solar Thermal Route.

(b) Solar Photovoltaic Route

A solar thermal device converts solar radiant energy into thermal energy. This is achieved by means of a black metallic surface (absorber) enclosed in an air-tight base and surrounded by thermal insulated materials from sides other than those exposed to sunlight. The exposed surfaces are covered with a transparent glazing of suitable materials such as glass or plastic. When exposed to solar radiation the blackened surface absorbs solar radiation and converts it into heat. The heat could be used for cooking of food, heating of water and air, evaporation



of liquids, generation of steam for heat applications and for power generation, drying of foodgrains, vegetables and fruit refrigerates etc.

### **Solar Cooker**

Solar cooker is a device which could be used for cooking of food during day time. Different types of solar cookers are available in the market. However, box type solar cooker is like a hot box, in which one can cook food without any cooking gas or kerosene, electricity, coal or fuel wood. This cooker works with the solar energy which is available free of cost. It, however, supplements the conventional fuel but does not replace it totally. In a solar cooker one can boil, bake and roast. A solar cooker works as a hot case-cum-cooker. The important parts of a hot box solar cooker include the (i) outer box, (ii) inner cooking box or tray, (iii) double glass lid, (iv) thermal insulator, (v) mirror and (vi) cooking containers.

### **Working Principle of solar cooker**

While in operation, the solar cooker is placed in the sun with its lid open. Direct sunrays enter the inner box through a double glass lid. A mirror placed on inner side of the lid reflects additional sunshine into the box. The black surface of the inner box and cooking containers absorb and trap the solar radiation. Direct and reflected solar radiation striking the cooking containers cook the food placed inside them.

### **Solar cookers are of two types :**

- (i) Box type and (ii) Parabolic Type.

### **Solar water heating systems**

Solar water heating systems use the flat-plate solar collectors with built-in channels or riser tubes attached to the absorber sheet. With a black paint coated absorber system the water can be heated up to a temperature of  $60^{\circ}$  to  $65^{\circ}\text{C}$ , while in selectively coated system the temperature of water can be raised to  $85^{\circ}$  –  $90^{\circ}\text{C}$ . For higher temperature evacuated tubular collectors are used. The flat-plate collector is basically a black surface that is placed at a convenient angle to the daily motion of the sun. It is provided with a transparent cover and appropriate insulation on all the sides. The heat transfer fluid is generally water, but air is also used.

### **Solar Dryers**

Solar drying systems have many applications. These can be used to dry vegetables, fruits, grains, fish, timber or any other perishable material. However, the technology for each type of material is different from the other. Therefore, the designs, mainly flow of air, temperature of hot air, etc are controlled using different control devices. Timber drying has been proved to be one of the two most useful technology of solar drying.

### **Solar desalination**

Solar desalination technology is used to convert impure water into potable water. When exposed to solar energy, the water starts evaporating and the water droplets get condensed



on the inner surface of the sloping roof and are collected with the help of channels fixed on the inner surface of the chamber. The purity of water is in the range of about 20 ppm which is good for drinking as well as for use in topping of batteries.

### **Solar Photovoltaic Technology**

Solar photovoltaic system directly and instantaneously converts sunlight into electricity through electronic processes. The basic building block is known as a solar cell. Solar cells are made of semiconducting materials; the most common material used being silicon. The solar cell operation is based on the ability of semiconductors to convert sunlight directly into electricity by exploiting the photovoltaic effect. In the conversion process, the incident energy of light creates mobile charged particles in the semiconductor which are then separated by the device structure and produce electrical current. When light falls over a silicon surface, part of it is absorbed in the silicon material and a voltage is generated. Typically, a silicon solar cell generates 0.6 volts and the current generated depends on the surface area of the solar cell. Therefore, to obtain sufficient current and voltage for running any electrical appliance, a number of solar cells are connected in series or parallel. To protect these solar cells from environment, a string of solar cells is placed on glass and is hermetically sealed. Such sealed strings of solar cells are known as solar photovoltaic modules. The photovoltaic modules are connected to form a solar photovoltaic array which can be used to energise any electrical load. Photovoltaic modules generate d.c. power. To operate electric appliances used in household suitable inverters are used to convert D.C. power into 220 volts 50 Hz A.C. power. Since the photovoltaic power is generated only under sunlight, storage batteries are used to preserve the photovoltaic power for use during night or whenever the sun is not shining. Grid interactive solar PV systems are also being used now-a-days which work without battery. However, this system pushes power to the grid only during day time.

### **Solar cell Technology**

Solar cells represent the fundamental power conversion unit of photovoltaic system. They are made from semiconductors and have much in common with other solid-state electronic devices such as diodes, transistors and integrated circuits. For practical operation solar cells are usually assembled into modules.

Many different solar cells are now available in the market and yet more are under development. The range of solar cells spans different materials and different structures in the quest to extract maximum power from the device while keeping the cost to a minimum. Devices with efficiency exceeding 30% have been demonstrated in the laboratory. The efficiency of commercial devices, however, is usually less than half this value.

Crystalline silicon cells hold the largest part of the market. To reduce the cost, these cells are now often made from multicrystalline material, rather than from the more expensive single crystals. Crystalline silicon cell technology is well established. The modules have a long lifetime (25 years) and their best production efficiency is approximately 18%.

Cheaper (but also less efficient) types of silicon cells made in the form of amorphous thin films are used to power a variety of consumer products. A variety of compound



semiconductors can also be used to manufacture thin film cells, for example, cadmium telluride on copper indium diselenide. These modules are now beginning to appear in the market and hold the promise of combining low cost with acceptable conversion efficiencies. However, the most widely used and technically developed type of solar cell is the silicon cell. Single crystal silicon of ultra-high purity is doped through its bulk with arsenic to produce n-type

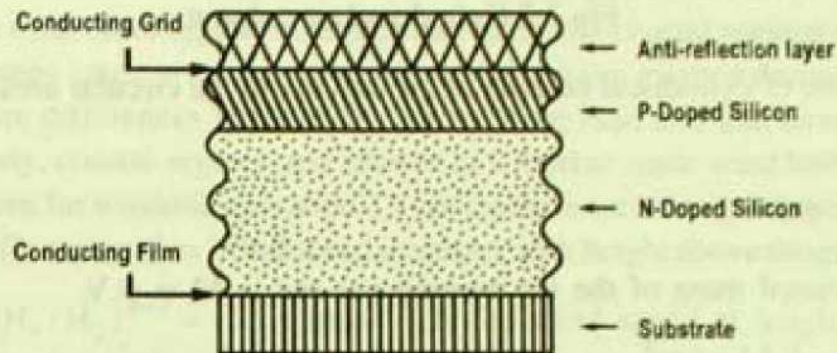


Fig.3.1

silicon. The substrate is subsequently doped with boron to produce p-type silicon. This type of cell is called a pn-junction solar cell.

### Application of solar photovoltaic cell

Once electricity is generated through a solar photovoltaic cell it can be utilised for any purpose like operation of pump, village electrification, running of small industries etc. The use of photovoltaic systems is increasing gradually throughout the world. In India many villages have been electrified with PV solar route. A large number of solar PV pumps are being used for irrigation and drinking water supply. Thousands of families in the rural areas are using solar lanterns and solar house lighting system.

### 3.3 Wind Energy

The energy in the wind has been used throughout the history in many parts of the world and was developed to such an extent that up to the time of the Industrial Revolution it was the most powerful and widely used energy source under human control. The fossil fuel burning heat engines that came with the industrial revolution led to the decline of wind power due to a better power/weight ratio and more predictable availability of power from the steam engine and the internal combustion engine. Today with greatly increasing oil prices there is a considerable revival in interest in wind power, both for water pumping and for electricity.

The energy available in the wind is not at all proportional to its speed, in fact, the output of a wind mill will vary with the cube of the windspeed. To have a rough estimate of available wind power, let us suppose a vertical blade of diameter  $D$  is rotating about a horizontal axis  $O$  at its centre, and a wind of velocity  $\phi$  is blowing horizontally towards the blade.



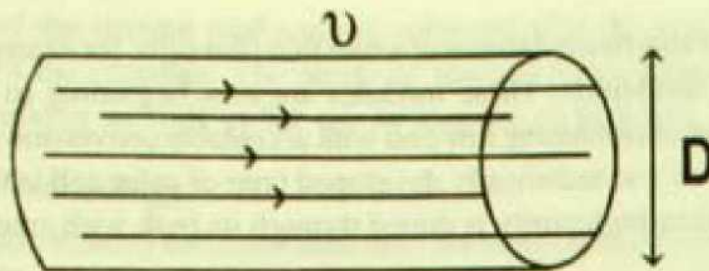


Fig.3.2 Cylindrical air column

The volume of cylindrical column of air moving to the circular area of rotating blade per sec

$$V = \frac{\pi D^2}{4} \cdot v.$$

The associated mass of the air moving per sec is  $M = \rho \cdot V$ ,

where  $\rho = 1.3 \text{ kg/m}^3$ .

Assuming that the velocity of the air is reduced to a very small value at the blade the available energy per sec would be

$$P = \frac{1}{2} m \cdot v^2 = \frac{1}{2} \rho \cdot V \cdot v^2$$

$$= \frac{1}{2} \cdot \frac{\pi D^2}{4} \cdot \rho \cdot v^3.$$

This available power is proportional to

i)  $v^3$ .

and ii)  $D^2$ .

But actual power is much less since the velocity of the air is not reduced to zero at the blade. In other words, doubling the wind speed has the effect of increasing the available energy by a factor of eight, while a halving of the windspeed reduces the energy to one-eighth. It is generally not practicable to make use of wind with speeds lower than about 5 to 7 mph (8 to 11 km/h) while wind with higher than about 30 mph (50 km/h) speed tends to be too powerful to be conveniently used. Therefore, most windmills are designed to make use of winds in the range mentioned; they do not function at lower windspeed and invariably either furl themselves or deliberately shed a lot of the available power at higher windspeed with the air in an automatic furling system to prevent any damage. Hence it is obviously important that there are reasonably frequent winds available at a proposed windmill site in the range normally above 10 mph or 16 kmph and certainly above 5 mph (8 kmph).

Although the wind is variable in an unpredictable manner from day to day the total wind energy passing over a particular site remains surprisingly constant if measured over



the years. Wind behavior is governed by a combination of global, regional, seasonal and diurnal weather patterns. Globally there are a series of wind belts of varying intensity and prevailing direction. There are mobile regional effects caused by hills or mountain ranges and the effect of lakes or sea shores. Seasonal changes such as monsoon etc are caused by the heating and cooling of a large continental land mass in the summer and winter.

In general the latitudes within  $10^\circ$  either side of the equator tend to have lower average wind speeds than most other regions. On the other hand, the coastal regions of hot countries or areas around large lakes in hot sunny regions tend to have marked diurnal winds caused by the temperature differences between the air over the land and that over the water. So, very approximately, coastal regions and islands in a marine trade wind belt often offer the best wind conditions for windmill operation. Of course there are some good sites for installation of windmills in hilly region also. Wind speed increases with height above the ground according to the formula  $(H_1/H_2)^{0.17} = \frac{V_1}{V_2}$  where  $V_1$  is the wind speed at height  $H_1$  and  $V_2$  at height  $H_2$ . The wind speed is generally recorded by weather stations at a standard height of 10 m above the ground — therefore if a windmill is placed 20 m above the ground it will feel a wind speed  $2^{0.17} = 1.125$  times that at 10 m, an increase of 12.5%. Obviously placing the windmill still higher will improve the wind power availability further, but this has to be paid for in terms of a more expensive tower. Most manufactures specify that the lower part of the rotor disc should be at least 6 m above the highest level of obstructions such as trees or buildings.

### Sizing of Windmills

There are, of course, two main applications for windmills:— pumping water and generating electricity. In general water pumping windmills have multibladed rotors, while most electricity generating machines have two or three bladed rotors more like an aircraft propellor in appearance. The reason for this is that a high starting torque is needed to get a water pump started, and the provision of many blades eases starting against a heavy load in light winds. Unfortunately a multibladed rotor is less efficient.

In order to arrive at reasonably accurate predictions of likely windmill performance in a given location regular wind records are needed, ideally for several years. The formula used to calculate power in the wind is

$$P = 0.00000323d^2v^3$$

where  $P$  = power in kW

$d$  = diameter in m

$v$  = velocity in kmph



Table 3.1 : Power generation in kW from electricity generating windmill with power co-efficient = 0.3

Wind Speed (kmph) \ Rotor diameter $\delta$ (m)	1.8	2.4	3.7	4.3	4.9
16	.043	.077	.173	.235	.307
24	.146	.259	.583	.794	1.04
32	.346	.614	1.38	1.88	2.46

### 3.4 Biomass

"Biomass" is a modern jargon for the oldest human energy resource. It means biologically derived material of any kind, all of which are potentially useful as a source of heat energy. Dried organic matter is invariably capable of being burnt and the heat produced in this process represents released solar energy stored chemically. The main forms of biomass in general use are of course wood and dried animal dung.

In general, wood and other vegetable matter normally have the following composition on moisture and ash free basis : 50% carbon, 6% hydrogen and 44% oxygen. The moisture content varies over a wide range from oven dry to about 90% on wet basis and ash content varies from 0.5 to 22%. Biomass, in particular, wood can be represented by the chemical formula  $CH_{1.44}O_{0.66}$ .

Though biomass conversion technologies and devices cover a very wide spectrum, these can be covered generally through four major categories.

- (i) Anaerobic digestion,
- (ii) Pyrolysis,
- (iii) Combustion and
- (iv) Gasification.

Anaerobic digestion has been extensively used in the country for high moisture organic materials, particularly dung. Many organic materials can be fermented in a sealed container (with air excluded). Under these conditions anaerobic bacteria can break down the biomass and yield in the process a feed gas (similar to naturally occurring 'marsh gas') consisting of about 60% methane and the rest containing mainly carbon dioxide. The following table indicates typical gas yields from various input materials (the gas in all cases being methane from 60 to 70 percent)



Table 3.2 : Gas yields from different biomass materials

Material	Gas yield per unit dry matter ( $\text{m}^3/\text{kg}$ )
Cow dung	0.1 – 0.3
Chicken dropping	0.3
Pig dung	0.4 – 0.5
Elephant grass	0.4 – 0.6
Sewage sludge	0.6

Major application has been for domestic and institutional cooking though a few installations for industrial heat as well as shaft power/electricity have also been made. The basic process is also used for treatment of industrial and municipal wastes and effluents with a number of installations for distillery effluent and human excreta having been made in various parts of the country. Being based on bacterial activity, the long retention times and high moisture/water content also result in relatively large plant sizes.

Pyrolysis could be largely used for production of various gaseous, liquid and solid fuels from biomass, with the most common example being conventional charcoal.

Biomass combustion can be said to be one of the oldest technologies known by mankind. Industrial revolution was largely based on combustion and biomass combustion has been widely practised all over the world for producing heat as well as for generation of shaft power/electricity.

Biomass gasification is basically conversion of solid biomass (i.e. wood/wood waste, aquacultural residue etc.) into a combustible gas mixture normally called "Producer gas" (or low Btu gas); the power is typically used for "woody" biomass and it involves partial combustion of such biomass. Partial combustion process occurs when air supply (or more precisely, oxygen) is less than adequate for combustion of biomass to be completed. Given that biomass contains carbon, hydrogen and oxygen molecules, complete combustion would produce carbon dioxide ( $\text{CO}_2$ ) and water vapour ( $\text{H}_2\text{O}$ ). Partial combustion produces carbon monoxide (CO) as well as hydrogen ( $\text{H}_2$ ) which are both combustible gases.

Solid biomass fuels are usually inconvenient, they have low efficiency of utilisation and can only be used for certain limited applications. Combustion is the normal conversion process, while direct thermal use in cooking, heating space and water, generation of steam is possible, generation of power, for example, requires high/medium pressure steam boiler along with steam engine or turbine with accessories. For small power needs this conversion technology is not only capital intensive and complex, it is also very inefficient.

### 3.5 Ocean Energy

Energy is stored by nature in the tides, waves, and thermal and salinity gradients of the oceans of the world.



The oceans receive, store, and dissipate energy through various physical processes. Energy exists in the form of tides, waves, temperature differences, salt gradients, and marine biomass.

### **Tidal Energy**

Tides are created by the gravitational attraction of the moon and the sun acting on the oceans of the rotating earth. The relative motions of these bodies cause the surface of the oceans to be raised and lowered periodically according to a number of interacting cycles. Tides in the open ocean have a maximum amplitude of about 1 meter, whereas tides closer to shore, such as those that occur in estuaries have substantially higher amplitudes.

Extraction of tidal energy is considered practical only when the tides are large and suitable sites for tidal plant constructions can be found. It is estimated that 50,000 MW of power could be generated from tidal power source in India. A modern tidal energy scheme consists of a barrage or dam that is constructed across an estuary and is equipped with a series of gated sluices to permit entry of water into the basin. The power is generated by using low head axial turbines which are large axial flow turbines having diameters as great as 9 m.

### **Wave Energy**

Ocean waves created by the interaction of winds with the sea surface, contain both kinetic energy, which is described by the velocity of the water particles and potential energy, which is a function of the amount of water displaced from the mean sea level. A wave energy device extracts energy from the sea and changes it to another form — usually mechanical motion or field pressure converting this energy to electricity, however, it is not simple because of the low frequency of the wave. Research and development activities are going on throughout the world on generation of electricity from wave.

### **Ocean Thermal Energy Conversion**

In the tropical and subtropical oceans of the world, a natural temperature difference exists between the surface water and those at depth. The concept of Ocean Thermal Energy Conversion (OTEC) exploits this temperature difference to drive power plants to produce electricity. Because the surface waters are warmed by the sun, OTEC can be considered an indirect solar technology. Unlike other solar technologies, however, a reliable OTEC plant would be able to generate electricity continuously because the temperature difference lasts 24 hours a day.

### **3.6 Geothermal Energy**

In geological terms geothermal energy is defined as the heat above the mean ambient temperature of the earth's solid core which is about  $8 \times 10^{30}$  Joules. The amount is enormous and represents 35 billion times the world's present total annual energy consumption. In reality however only a tiny fraction of natural heat can be extracted from the earth's crust, mainly for economic reasons, which limits exploitation to a maximum depth of 5 km. To this depth the temperature of the crust increases at an average rate of 30 to 35°C per km.



## Chapter 4

### Vacuum Techniques

#### 4.1 Introduction

The word vacuum is derived from the Greek word meaning 'empty'. Vacuum plays a basic and indispensable role in present day technology and is used by a wide variety of scientists, chemists, biologists and engineers who work in research, development and industrial production.

The SI unit of pressure is Pascal (Pa) which is equivalent to one Newton/m<sup>2</sup>. The other units mbar and Torr are also commonly used. The relationship between these units is

$$100 \text{ Pa} = 1 \text{ mbar} = 0.76 \text{ Torr.}$$

It is useful to divide the total pressure range into four regions, namely,

low vacuum— atmospheric pressure to  $10^2$  Pa

medium vacuum—  $10^2$  Pa to  $10^{-1}$  Pa

high vacuum—  $10^{-1}$  Pa to  $10^{-5}$  Pa

ultra-high vacuum—  $10^{-5}$  Pa and less

In the low vacuum region, the main property is the force exerted by the atmosphere and this is used for mechanical handling, vacuum forming, vacuum brakes, degassing of fluids.

The applications in the medium vacuum range are extensive and include processes such as vacuum drying and vacuum freeze drying for the food and pharmaceutical industries and vacuum distillation for the chemical industry. In many of these processes an important factor that has to be considered is the vapour pressure of the fluid which, of course, is often water and the effect of this on the vacuum pumps must also be taken into account. It is necessary that the pressure in the system should be less than the saturated vapour pressure of the fluid at the appropriate temperature. Thus in vacuum drying at room temperature the pressure must be less than about 100 Pa, whereas for vacuum freeze drying in the range  $-50^\circ\text{C}$  to  $-180^\circ\text{C}$ , the pressure must be in the region of 1 Pa.

The high vacuum region has many applications and includes the production of special materials for the metallurgical, electronics and aircraft industries and other processes such



as electron beam welding. In TV, X-ray and gas discharge tubes, electron microscopy and particle accelerators it is necessary to use high vacuum but perhaps the most important process in this pressure range is that of vacuum evaporation of thin films for lens blooming and many aspects of the semiconductor and computer high technology industries. Nearly all of these applications demand that the "mean free path"—the average distance travelled by the gas molecules between collisions—must be greater than the dimensions of the vacuum chamber. The approximate value of the mean free path in terms of the pressure,  $p$  pascal is given by the equation

$$\text{mean free path} = 0.66/p \text{ cm}$$

Thus at a typical high vacuum pressure of  $10^{-3}$  Pa the mean free path is about 6.6 m which is greater than the dimensions of most high vacuum chambers.

The pressure in the atmosphere is about  $10^{-3}$  Pa at a height of  $10^6$  m so that some space simulators require ultra-high vacuum. On the other hand research in thermonuclear fusion uses ultra-high vacuum in order to achieve extremely high gas purities. However in this pressure range the mean free path is very large and it is more important to consider the molecule-surface collisions rather than the molecule-molecule collisions. If we assume that at a pressure of  $p$  Pa all gas molecules arriving at a surface remain on that surface, then the time  $t$  seconds to form a monolayer of, for example, nitrogen molecules, on that surface is given approximately by the equation

$$t = 3 \times 10^{-4}/p$$

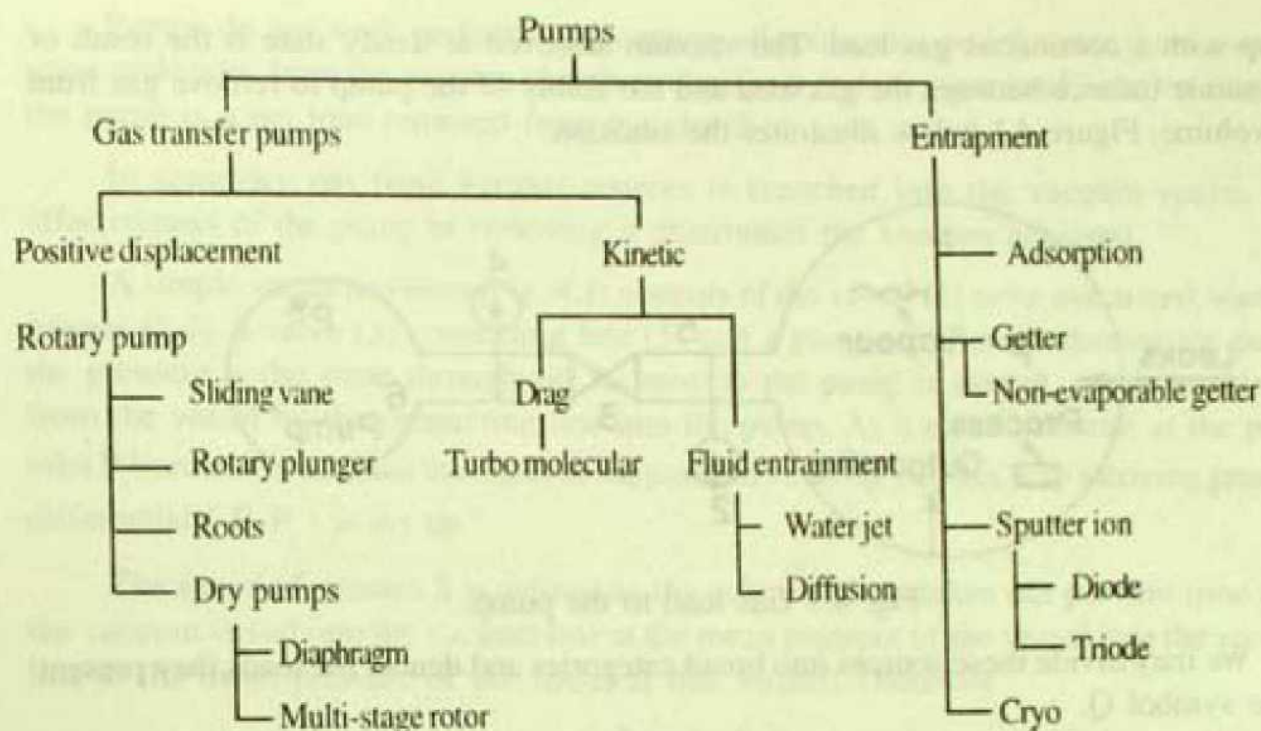
Thus at  $10^{-4}$  Pa the time is a few seconds whereas at  $10^{-8}$  Pa the time is several hours and this makes it possible to conduct measurements on atomically clean surfaces. Ultra-high vacuum is therefore an essential requirement in all surface studies including field-electron and field-ion microscopy (FEM and FIM), Auger electron spectroscopy (AES), secondary ion mass spectroscopy (SIMS) and photoelectron spectroscopy (PES).

However, since about 1985 there has been an increasing demand to achieve pressures less than  $10^{-9}$  Pa in a number of specialised areas. One example is the "CRYRING" or Electron Storage Ring (ESR) heavy ion accelerator which requires pressures less than  $10^{-10}$  Pa in the storage mode.

### Vacuum pumps

A vacuum pump is a device of creating, improving and/or maintaining a vacuum. Two basic categories exist, the gas transfer and the entrapment groups. The gas transfer group can be subdivided into positive displacement pumps in which repeated volumes of gas are transferred from the inlet to the outlet usually with some compression, and the kinetic pumps in which momentum is imparted to the gas molecules so that gas is continuously transferred from the inlet to the outlet. In contrast, entrapment pumps are those which retain molecules by sorption or condensation on internal surface. Accordingly, the vacuum pumps are classified as follows :





### Ranges of operation of vacuum pumps

#### Range

$10^5$  to  $10^2$  Pa : Pumps which operate in this low vacuum region are some of the positive displacement pumps, ejector pumps, adsorption pumps and cryopumps.

$10^2$  to  $10^{-1}$  Pa : Pumps which operate in this medium vacuum region are some positive displacement pumps, ejector pumps and adsorption pumps.

$10^{-1}$  to  $10^{-5}$  and  $10^{-5}$  and beyond : High vacuum pumps and ultra high vacuum pumps include molecular diffusion oil vapour pumps, diffusion mercury-vapour pumps, turbomolecular pumps, sorption pumps (including plain getter pumps, arc discharge getter pumps, getter-ion pumps, sputter-ion pumps) and cryogenic pumps.

In the medium and high vacuum ranges, the removal of condensing vapour is speeded up by the use of cold trap usually cooled to the temperature of liquid nitrogen.

In choosing a vacuum pump we must also consider the economy, power consumption, requirements for cooling operating fluid, rotational speed level of noise and vibration etc.

### 4.2 Qualitative description of the pumping process :

To achieve the required vacuum is not simply a matter of removing a sufficient quantity of the air originally in the vessel. This indeed has to be removed but we then find that there are continuous sources which launch gas into the volume and which present the

pump with a continuous gas load. The vacuum achieved at steady state is the result of a dynamic balance between the gas load and the ability of the pump to remove gas from the volume. Figure 4.1 below illustrates the situation.

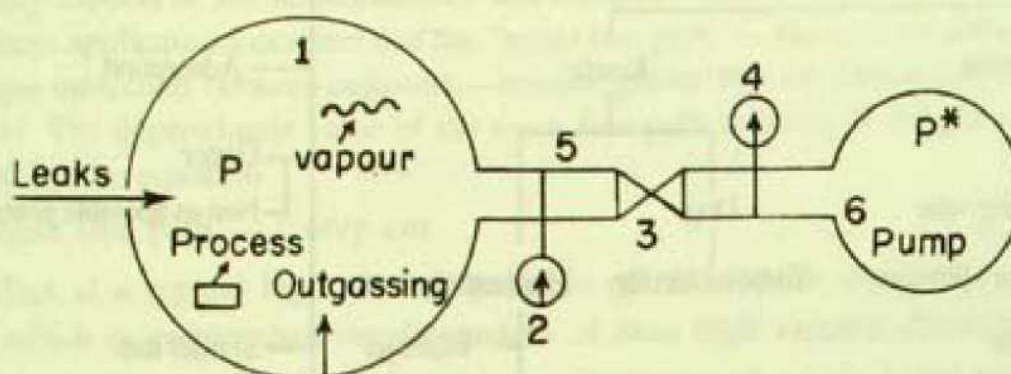


Fig. 4.1 Gas load to the pump.

We may divide these sources into broad categories and denote the loads they present by the symbol  $Q$ .

(a) **Leaks,  $Q_L$**

These may be real leaks due to passageways through the vacuum wall from outside the vessel or virtual leaks due to gas being trapped in localities from which it can emerge only slowly into the vacuum surroundings.

(b) **Vaporisation,  $Q_V$**

Materials which exert a significant vapour pressure are present in a vessel. They contribute a gas load. Water vapour from imperfectly dried components has to be strictly avoided in HV and UHV applications.

(c) **Outgassing,  $Q_G$**

This term describes the release of gas from the internal surface of the vacuum wall and the surfaces of components inside the vessel. It forms the principal source of gas in many systems and limits the degree of vacuum which can be achieved.

(d) **Process generated gas,  $Q_p$**

Many processes carried out in vacuum cause the release of gas, often from materials which are heated. For example, in vacuum degassing applications metals are heated to high temperature to get rid of dissolved gas.

(e) **Others**

Depending on the type of pump used in a given application, there may be a tendency for the vapour of lubricants or, in the case of a diffusion pump, vapour of the working fluid, to "back-stream" into the vacuum chamber. Precautionary measures such as interposed traps may be used to reduce this.



Pumps do not work perfectly and capture all molecules which enter them, so that some molecules from the vacuum chamber return to it and the "gas load" shown entering the pump is a net load removed from the chamber.

In summary gas from various sources is launched into the vacuum space. The effectiveness of the pump in removing it determines the vacuum achieved.

A simple vacuum system (Fig. 4.1) consists of the vessel (1) to be evacuated, vacuum gauges (2,4), a valve (3) connecting line (5) and a pump (6). Before starting the pump, the pressure is the same throughout. As soon as the pump is started, gas is transferred from the vessel by the connecting line into the pump. As a result pressure at the pump inlet  $P_i$  becomes lower than the outlet of the pumped vessel (P). In this way a driving pressure differential ( $P - P_i$ ) is set up.

The speed of exhaust S is defined as the volume of gas taken out per unit time from the vacuum vessel into the vacuum line at the mean pressure of the vessel into the vacuum line at the mean pressure of the vessel at that instant, Therefore

$$S = dv/dt$$

Since pressure is constant across any particular cross section, we define 'throughput' Q by

$$Q = p.dv/dt = p.S.$$

Q is the basic quantity which specifies the gas flow,

so,  $S = Q/P$

Since it assumes a special role, we denote  $S^*$  as the speed at the inlet of a pump. Conductance C, measures the ease of flow and is defined by

$$C = Q/(p_1 - p_2)$$

where ( $p_1 - p_2$ ) is the pressure difference between two regions and Q is throughput.

Components may be connected in series or in parallel. The effective conductance is

for C's in parallel

$$C = C_1 + C_2 + C_3 + \dots$$

$$Q = Q_1 + Q_2 = (C_1 + C_2) (P_1 - P_2)$$

for C's in series

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots$$

From figure 4.2 we see, the throughput Q is

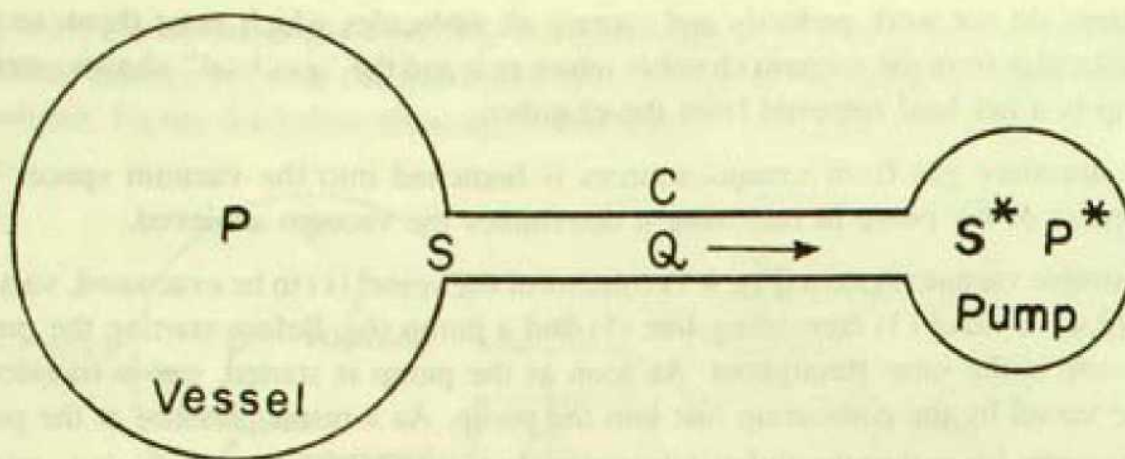


Fig. 4.2 Dependence of pumping speed on conductance.

$$Q = C(P - P^*) = SP = S^*P^*$$

$$P = \frac{Q}{S} \text{ and } P^* = \frac{Q}{S^*}$$

$$Q = C\left(\frac{Q}{S} - \frac{Q}{S^*}\right)$$

$$\text{which give } S = S^* \left[ \frac{C}{C + S^*} \right]$$

This equation shows that the effect of conductance is always to reduce the speed at the vessel. This equation is the fundamental equation of the vacuum system.

Pumps for the production of high vacuum may be roughly divided into two categories:—  
(a) those which pump air from a vessel at atmospheric pressure, these are usually mechanical pumps and (b) those which need a fore-vacuum or pumps which begin to operate below a certain limiting pressure. Pumps classified under (a) are commonly called "backing" pumps. We will describe Cenco—Hyvac pump in this category.

### 4.3 Rotary oil pump :

This is a favourite pump in laboratories and lamp factories. A rotor revolves inside an outer cylinder, but the rotor has a single spring operated moving vane, and the inlet and outlet ports are close together.

Referring to figure 4.3 there is an inner fineworked steel cylinder. This rotates eccentrically about the shaft B inside the steel cylindrical casing C. The vane is a spring operated by the arm D. The arrow on the rotor indicates the direction of movement. The air, in contact with the vessel being evacuated via port E and occupying volume V, is forced by the movement of the rotor and vane into a smaller volume near the outlet valve F, and so into the atmosphere. The whole mechanism is immersed in the oil in the box G. The commercially made "Hyvac" consists of two such units mounted side by side on a common motor-driven shaft; the two pumps are in series.



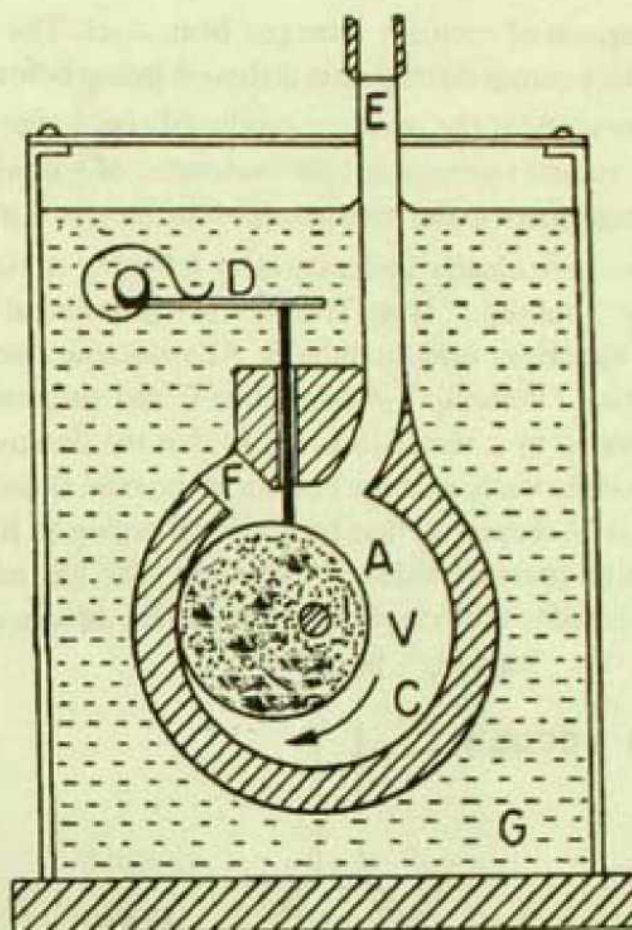


Fig. 4.3 Cenco Hyvac Rotary Oil Pump.

In the first stage the gas is admitted via E to the crescent-shaped space. In the second stage, the gas is admitted behind the rotor. Further compression follows in the next stage and finally the valve F opens and the gas is expelled. The speed of pumping is about 6 liters a minute and the vacuum attainable is about  $10^{-3}$  mm of Hg.

When using oil-filled "backing" pumps, care must be taken to ensure that, if the pump is left stationary over a long period, air is not sucked back from the pump into any evacuated vessel connected to it. To prevent this, the pump is either fitted with a self-sealing oil-pump, or a reservoir is fitted above the pump or the tap arrangement to allow the vessel to remain under vacuum whilst the oil-pump is at atmospheric pressure.

The mechanical oil-pumps are incapable of pumping condensable vapours such as water vapour, oil-vapour, tap-grease vapour, etc. It is advisable to protect such pumps by a small boat of phosphorus pentoxide placed conveniently between the pump and the vessel. This prevents water vapour spoiling the pressure, and also harming the precision-worked surfaces of the "backing" pumps' mechanism.

#### 4.4 Mercury Diffusion Pump :

The principle involved in this type is that in a mixture of gases, the diffusion of a gas occurs from a region where its partial pressure is higher to that where it is lower, irrespective of the total pressure in the two regions.



In this pump a stream of mercury emerges from a jet. The vessel being exhausted is first pumped by the fore-pump through this diffusion pump before the mercury is heated. Then on boiling this mercury at the pressure produced (i.e. lower than  $10^{-2}$  mm. Hg) the high-velocity mercury vapour stream traps the molecules of gas which diffuse into it from the vessel and guides them down to the fore-pump where the gas is passed to the atmosphere.

The action will be more clearly understood by referring to figure 4.4. This illustrates a diffusion pump of the "umbrella" type. The mercury is boiled at the "backing" pressure in the reservoir A and vaporizes into the tube B. At a suitable rate of heating, this stream impinges on the reflecting "umbrella" shaped cone C and streams down the annular tube D. The pump is surrounded by a water-jacket E so that the downward stream of mercury finally condenses against the walls near the bottom of the tube D and returns to the reservoir by tube F. This tube F is U-shaped so that mercury collecting in it prevents direct contact between the low- and high-pressure sides of the pump. The gas molecules in the space J, which are directly connected to the vessel being exhausted, diffuse into this mercury stream at D and are forced to the "backing" pump.

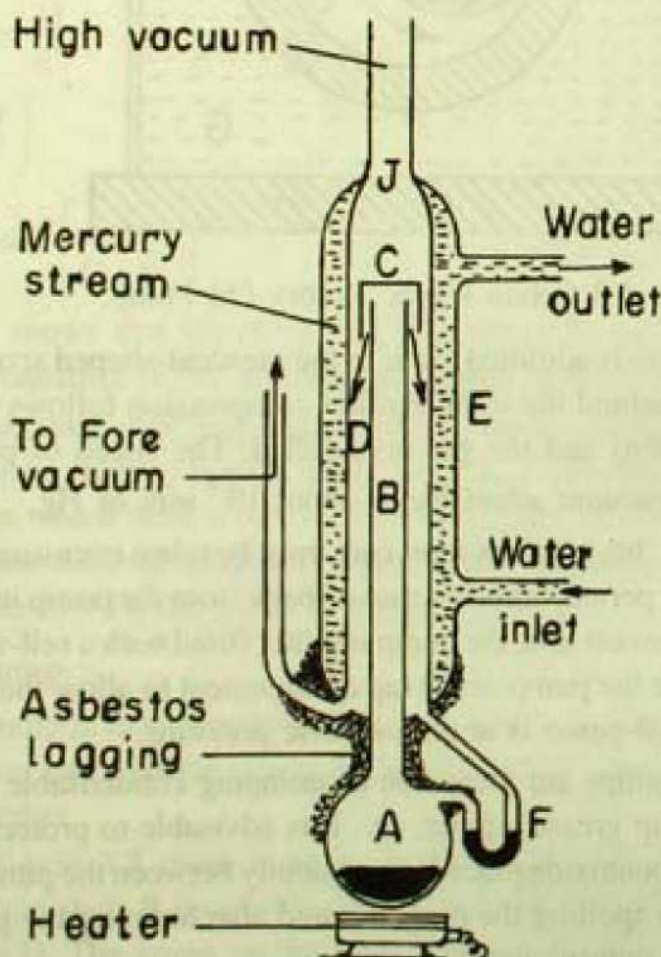


Fig. 4.4 Mercury Diffusion Pump.

Vacuum of less than  $10^{-6}$  mm. may be obtained, but while theoretically there is no limit to the vacuum attainable, in practice two disadvantages arise: (a) the exhaust speed is slow, (b) careful regulation of the temperature of the mercury vapour is required.



By a modification of the diffusion pump, by which it becomes a condensation pump, Langmuir eliminated both these disadvantages. The main advantage consists in cooling the mercury vapour thoroughly at the jet, so that condensation occurs, back diffusion of the mercury vapour being thereby completely eliminated. The apparatus has the advantage of requiring no critical conditions and the size of the orifice may vary over a wide range. The pump is very efficient, reducing a pressure of 1 mm to  $10^{-5}$  mm of Hg. in 80 sec. with a speed of working of nearly 4 litres/sec. At a slow heating rate it will require a low 'backing' pressure, while too fast heating process will send the mercury vapour undesirably into upper parts of the pump. To apply the condensation principle satisfactorily, the end of the diffusion nozzle must be inside the water jacket and the length of the pump should be great enough to prevent any appreciable quantity of gas diffusing back against the mercury. The width of the slit equals the mean free path at the maximum pumping speed. Back streaming of pump oil vapours is a serious problem because it adds impurities. Traps and baffles are used to restrain vapour molecules from passing into the vacuum chamber without the flow of the pumped gas in the opposite direction. The ultimate pressure obtained with diffusion pumps is limited by the gas emission from various parts of the vacuum system.

Many varieties and sizes of commercial pump are now available which operate on the above principles, but in most cases the mercury has been replaced by special oils as the working substance. Robust vessels of copper and brass contain the oil which is vaporized by electrical heating. High efficiency is ensured since the vapour pressure of these oils is extremely low at ordinary temperatures. The possibilities of metallic corrosion in various parts of the vacuum system are negligible compared with those when mercury is used. The high molecular weight of oil compared with that of mercury also ensures greater pumping speeds, sometimes as high as 3500 litres/sec.; but the lower temperature essential for oil pumps to avoid decomposition of the oil requires the use of larger and more expensive backing pumps to give a greater initial backing vacuum.

#### 4.5 Measurement of high vacuum :

A pressure gauge is a device to measure the pressure produced by a pump. According to the principle of operation, vacuum gauges may be classified as follows :

##### 1. Total pressure gauges :

- a) Hydrostatic pressure gauge — depends on the isothermal compression of an ideal gas — McLeod gauge (range  $10^3$ – $10^{-2}$  Pa)
- b) Mechanical gauges — elastic deformable element is used as sensor and the deformation is a measure of the vacuum — Bourdon Tube, Diaphragm etc. (Range  $10^5$ –1 Pa)
- c) Liquid-level vacuum gauges — Liquid manometers (Range  $10^5$ – $10$  Pa)
- d) Thermal conductivity gauges — dependence of thermal conductivity of gases on pressures is used — thermocouple and Pirani (Range  $10^4$ –1 Pa).



- e) Ionisation gauges — ionisation of gases is utilised. (Range  $10^2$ – $10^{-10}$  Pa).  
 There are two types :
- i) Hot cathode gauges — electrons produced by hot filament are accelerated by an electric field — Schulz and Phelps, Orbitron, Magnetron.
  - ii) Cold cathode gauges : glow discharge in electric and magnetic fields (Penning discharge) is used to increase the path length of electrons — Penning, Magnetron, Inverted Magnetron.

## 2. Partial pressure gauges :

Ions are separated according to mass by electromagnetic means. So that a mass spectrum is produced.

We will consider only three important gauges used frequently in the laboratory to measure the low pressure produced by exhaust pump.

### McLeod Gauge :

It was developed by McLeod in 1874. To put the measurement of the vacuum on a good quantitative basis, the McLeod gauge is used. Figures 4.5 (a) and (b) show two convenient forms of the gauge.

A pyrex glass bulb A is attached to a fine capillary tube B. C is a side-arm provided with a capillary at D which has the same bore as the B capillary, and runs close to and parallel with it. The vacuum to be measured is connected to the gauge via the tube E. The glass system is first thoroughly cleaned and filled with mercury up to the point marked H in the figure. In Figure 4.5 (a) this mercury level can be raised by lifting bulb G so that if the gauge is evacuated then it is completely flooded with the mercury. In Figure 4.5 (b) the top of the gauge is less than 760 mm. above the mercury level in the reservoir, and so the gauge will be normally filled with mercury if it is pumped free of air. To draw it out of the gauge, the mercury is sucked down by pumping the reservoir at G through tube J with a "backing" pump. To permit the mercury to rise again, an inlet into the reservoir is provided at S. This air-inlet can be turned on or off by means of the tap T.

To read the pressure the mercury is allowed to rise. A sample of the gas to be measured is trapped by the rising mercury in the volume between the marks H and I on the gauge. H is the point where the side-arm C is joined to the main gauge. Suppose this volume, i.e. the volume of bulb plus capillary, is V. The mercury goes on rising until the level in the side-arm C is opposite the top end of capillary B. Then the rise is stopped, either by bringing the reservoir bulb to rest or by closing the tap T of figure 4.5 (b). The mercury rise will compress our sample volume of gas into the capillary so that it occupies some length h. The cross-section a of the capillary is measured before setting up the gauge. Then the gas sample has been compressed from volume V to volume ah, and the head of mercury which is compressing it to this volume is also h, a is the cross section. Applying Boyle's law for the perfect gas



$$P.V. = ah.h$$

where P is the pressure of the gas which is required, and h is in mm. of mercury; therefore,

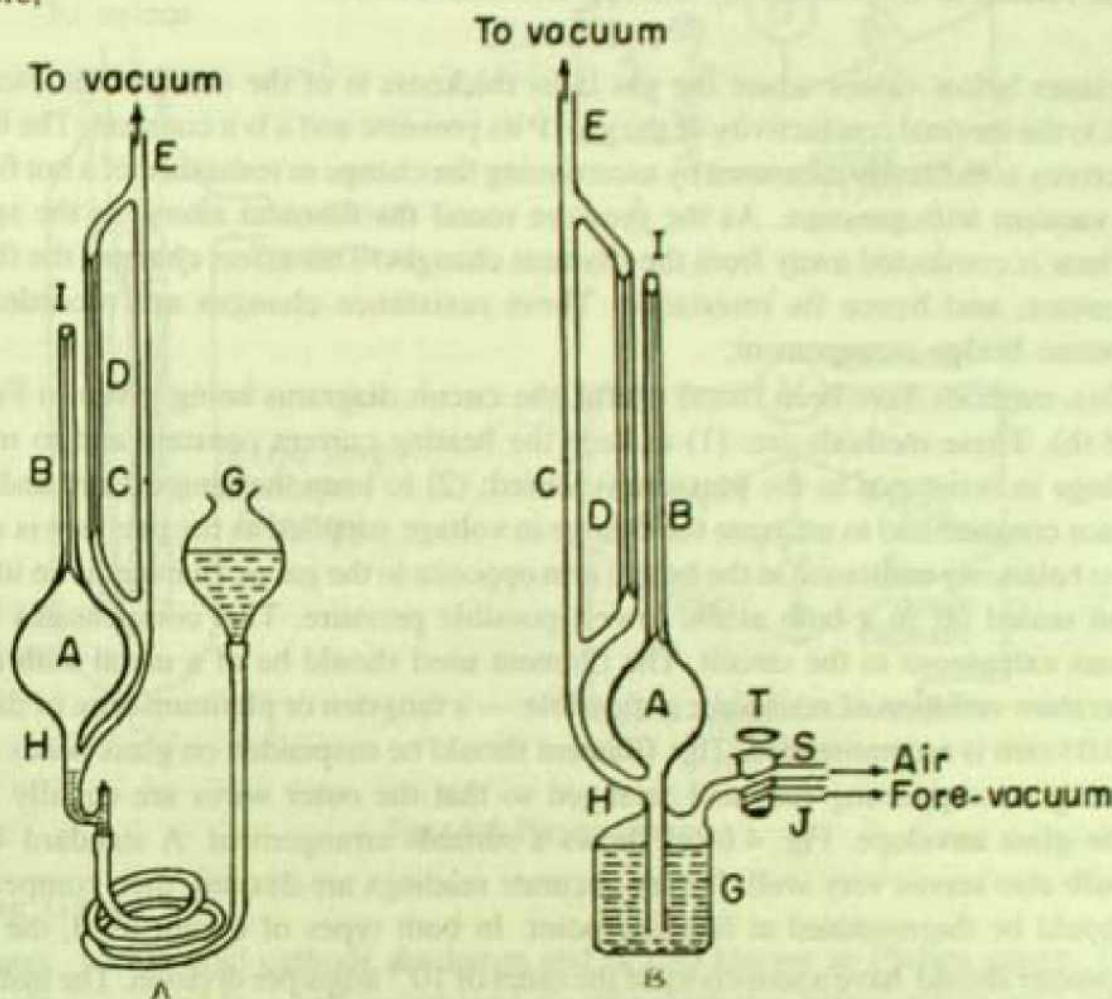


Fig. 4.5 McLeod Gauge.

$$P = \frac{ah^2}{V}$$

a and V are constants in this relation, and so the pressure can be measured in mm. of mercury on a square law scale.

If the bulb A is about 300 cc and the diameter of the capillary tube 1 mm then a pressure of  $10^{-5}$  mm Hg can be readily measured.

The McLeod gauge is the most used gauge in vacuum practice, and all the other gauges described below are calibrated against it. Its chief advantage is that it gives, in a straightforward manner, an absolute measurement of the vacuum. It suffers from one great disadvantage : it will not accurately register the pressure due to condensable gases and vapours. A second minor trouble is that the gauge must, inevitably, be connected to the system by a fair length of narrow tubing and so it usually takes about one minute for the gauge and system pressures to equalize.



## Pirani Gauge

The pirani Gauge is based on the physical principle that the thermal conductivity of a gas is related to its pressure by the approximate relation:

$$K = a.P.$$

for pressure below values where the gas layer thickness is of the order of the mean free path.  $K$  is the thermal conductivity of the gas,  $P$  its pressure and  $a$  is a constant. The thermal conductivity is indirectly measured by ascertaining the change in resistance of a hot filament in the vacuum with pressure. As the pressure round the filament alters, so the speed at which heat is conducted away from the filament changes. This effect changes the filament temperature, and hence its resistance. These resistance changes are recorded by a Wheatstone bridge arrangement.

Two methods have been found useful, the circuit diagrams being given in Figs. 4.6 (a) and (b). These methods are: (1) to keep the heating current constant and to measure the change in resistance as the pressure is varied; (2) to keep the temperature and so the resistance constant and to measure the change in voltage supplied as the pressure is altered. The best balancing resistance in the bridge arm opposite to the gauge filament is an identical filament sealed off in a bulb at the lowest possible pressure. This compensates for the variations extraneous to the circuit. The filament used should be of a metal with as high a temperature variation of resistance as possible:— a tungsten or platinum wire of diameter about 0.03 mm is recommended. This filament should be suspended on glass beads around a central glass supporting rod, and arranged so that the outer wires are equally spaced from the glass envelope. Fig. 4.6 (c) shows a suitable arrangement. A standard 40 watt lamp bulb also serves very well. If very accurate readings are desired, then compensating bulb should be thermostated at freezing-point. In both types of circuit used, the bridge galvanometer should have a sensitivity of the order of  $10^{-8}$  amps per division. The instrument has to be calibrated against a McLeod gauge.

Using the second method, due to N.R. Campbell, if  $V$  is the potential for a pressure  $P$  in the gauge and  $V_0$  is the potential for the lowest possible pressure, then

$$\frac{V^2 - V_0^2}{V_0^2} = k.P.$$

The useful range of the Pirani gauge is from  $10^{-1}$  mm. Hg to  $10^{-4}$  mm.Hg.

The chief drawbacks of the gauge are that (i) it is quite unsuitable for use with organic vapours, for they 'poison' the filament, (ii) it is not an absolute gauge and has to be calibrated against a McLeod or any other absolute gauge, (iii) it is not suitable for measurement of pressures below  $10^{-3}$  mm, because heat loss occurs mostly by radiation rather than by conduction, (iv) in the range  $10^{-3}$  to  $10^{-5}$  mm, it requires some manual adjustment, which obviously cannot be made reliable. These drawbacks were removed by Scott, in 1939, by including a triode valve in the circuit of the gauge, making its working both smooth and automatic.



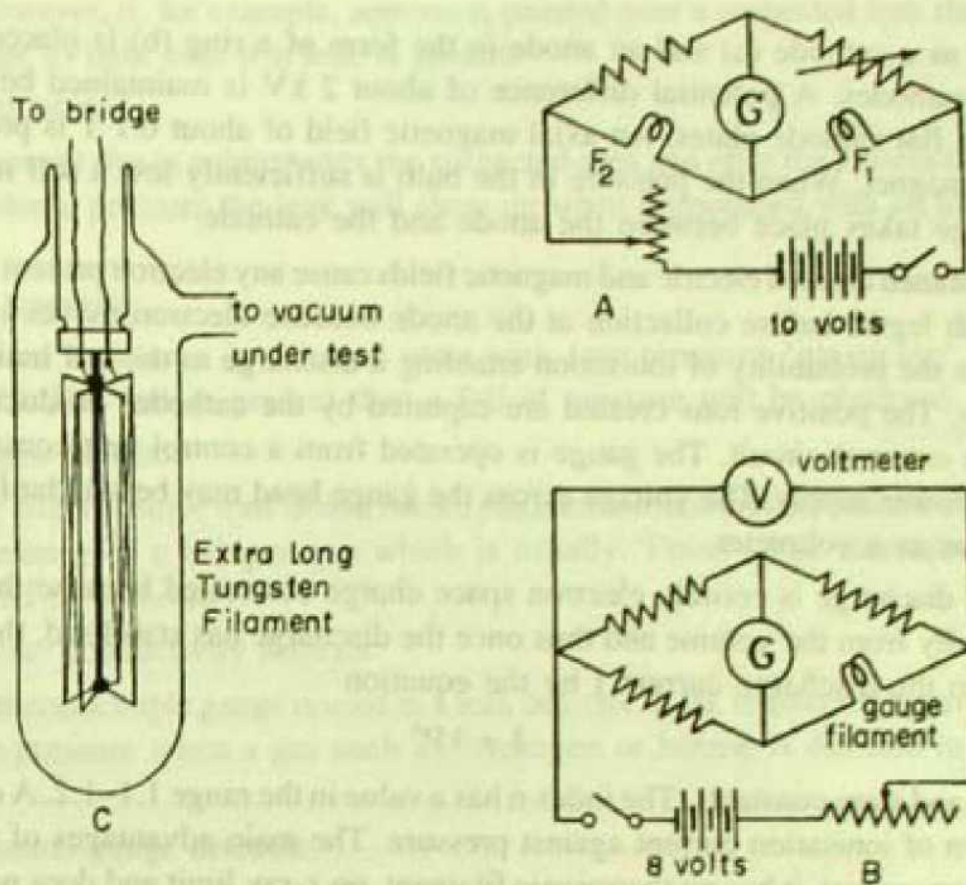


Fig. 4.6 Pirani Gauge.

### Penning Gauge :

It operates with a cold cathode discharge and is also known as Philips gauge. The gas in the bulb is ionised by a self-maintained electric discharge between two cold cathodes. The gauge is illustrated in the figure 4.7. The bulb is enclosed by a metal envelope

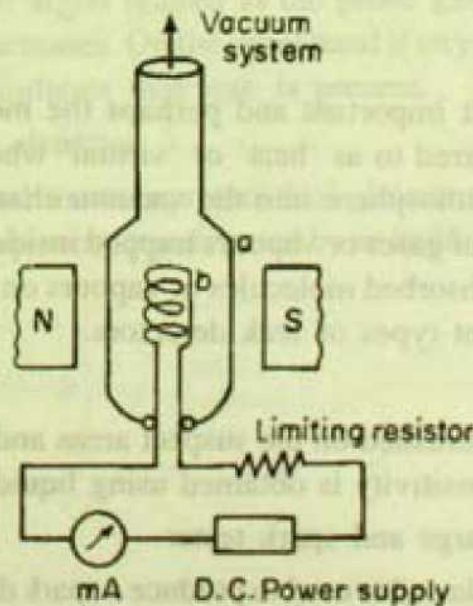


Fig. 4.7 Penning Gauge

which works as a cathode (a) and an anode in the form of a ring (b) is placed midway between the cathodes. A potential difference of about 2 kV is maintained between the anode and the flat cathode plates. An axial magnetic field of about 0.1 T is provided by a permanent magnet. When the pressure in the bulb is sufficiently low a self maintained glow discharge takes place between the anode and the cathode.

The combined crossed electric and magnetic fields cause any electron present to achieve very long path lengths before collection at the anode because electron moves in a spiral. This increases the probability of ionisation enabling a discharge to be self maintained at low pressures. The positive ions created are captured by the cathodes producing an ion current in the external circuit. The gauge is operated from a control unit consisting of a rectified AC power supply. The voltage across the gauge head may be standardized using a milliammeter as a voltmeter.

The gas discharge is entirely electron space charge controlled because the ions are extracted rapidly from the volume and thus once the discharge has stabilised, the pressure  $P$  is related to the discharge current  $I$  by the equation

$$I = kP^n$$

where  $k$  and  $n$  are constants. The index  $n$  has a value in the range 1.1–1.2. A calibration curve is drawn of ionisation current against pressure. The main advantages of this gauge are that it is very robust, it has no thermionic filament, no x-ray limit and does not produce any thermal radiation. However it is normally considered to be less accurate than the thermionic gauge because it is not directly proportional to  $P$ . It also has a relatively large pumping speed,  $10^{-2} - 1 \text{ S}^{-1}$ , and sometimes there are problems with striking the discharge at low pressures. Nevertheless it is easy to operate and is very popular for many scientific and industrial processes in the pressure range  $1-10^{-5} \text{ Pa}$ ; it can even be extended to  $10^{-7} \text{ Pa}$  by use of a current amplifier.

#### 4.6 Leaks

Leak detection is the most important and perhaps the most tedious aspect of vacuum technology. Leaks are referred to as 'heat' or 'virtual' where in the former case the gas passes from the external atmosphere into the vacuum chamber and in the latter it arises either due to the evolution of gases or vapours trapped inside the vacuum envelope in holes or channels, or due to the absorbed molecules or vapours on the inside walls of the vacuum system. There are different types of leak detectors.

##### a) Soap Bubbles

The soap solution is brushed on the suspect areas and bubbles will be seen if a leak is present. Higher sensitivity is obtained using liquids of lower surface tension.

##### b) Electrical discharge and spark tester

A high frequency Tesla coil is used to produce a spark discharge which will concentrate over a leak in an evacuated glass vessel. Alternatively an electrical discharge tube can be positioned above the rotary pump. The discharge would normally have a pinkish



colour. However, if, for example, acetone is painted over a suspected leak the colour will change to light blue if a leak is present.

c) Penetrant dye

The fluorescent dye is painted over the suspected area and after the system is opened to atmospheric pressure the leak will show up when illuminated with an ultraviolet lamp.

d) Leak covering

A suspected leak area is covered over with low pressure 'plasticine' such as Apiezon Q. If a leak is present then a fall of pressure will be observed.

e) Halogen detector

The probe emits positive ions from a heated platinum surface and the emission increases in the presence of a halogen gas which is usually 'Freon'. The increased current is indicated by a meter.

f) Thermal conductivity detector

Pirani or thermocouple gauge is used as a leak detector. Leak is detected by an apparent change in pressure when a gas such as, hydrogen or butane is directed in the area of leak.

g) Ionisation gauge detector

If a probe gas such as helium is directed towards the suspected leak area or if acetone is painted on this area, there will be change in the gauge sensitivity and an apparent change of pressure is indicated due to the presence of a leak.

h) Ion pump detector

Ion pump can be used as a leak detector by noting the change of pumping speed for different gasses. For example, the speed for argon is three or four times less than that of oxygen. Hence if argon is used as the probe gas, the gauge reading of the ion pump power supply increases. On the other hand if oxygen is used, the reading decreases. These two tests indicate that leak is present.

i) Mass spectrometer leak detector

This is the most sensitive and most important leak detector and at the same time it is very expensive. If a leak is present, it is detected by an audio signal or on electronics panel.

## SECTION-II

### Electronics

## Chapter 5

### Feedback

#### 5.1 Introduction

Feedback plays a very important role in almost all electronic circuits—both analogue and digital. The application of feedback makes it possible in some cases to improve the performance of the circuit involved to a marked degree. The use of feedback, in an improperly designed system, however, can produce a system with worse characteristics.

The principle of feedback is that we sample the output of the electronic system and compare the sample with the input signal. A difference between the output sample and the input creates an *error* signal, which is fed back into the input to reduce the error or to drive the output toward correspondence with input. This reverse transmission of signal, deliberately introduced and controlled, is an inherent characteristic of a feedback system.

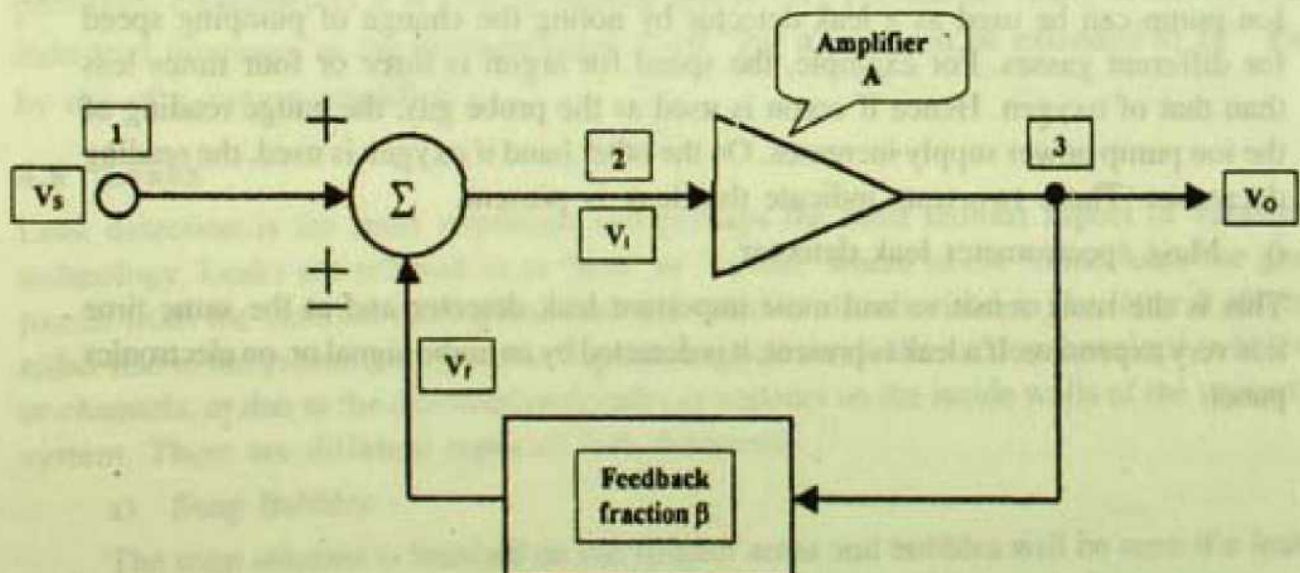


Fig. 5.1 Block diagram of feedback circuit



## 5.2 Basic Principles

We first show the effect of feedback on amplifier gain. An amplifier will be viewed as a "black box", represented by the triangular symbol of Fig 5.1. No knowledge of the internal circuit is required, except that it has an overall gain  $A$ . The output signal is sampled by the  $\beta$  network, giving an output  $V_f$ , called the feedback voltage. The feedback sample  $V_f$  is added to the source input  $V_s$  at the summing point  $\Sigma$ , to produce the *error* signal  $V_e$  at the input port of the actual amplifier.

The introduction of the  $\beta$  network requires:

- i) That the  $\beta$  network does not load the output or input circuits of the amplifier.
- ii) That the input signal be transmitted forward through the amplifier only. In practice, forward transmission through the amplifier is small.
- iii) That the reverse signal be transmitted from the output to the input only through the  $\beta$  network.

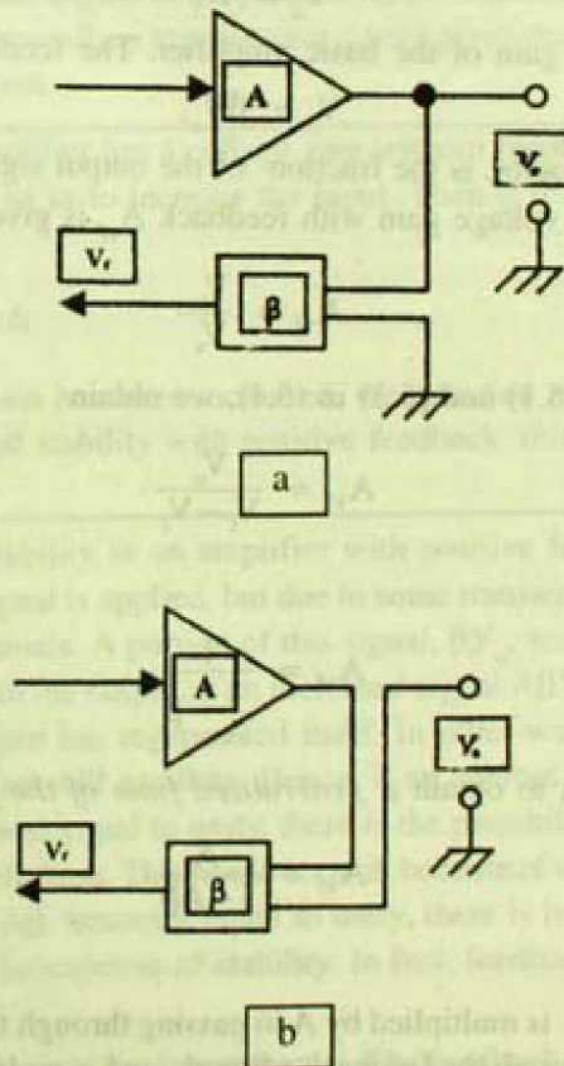


Fig. 5.2. Output sampling: (a) voltage, (b) current.

## Sampling network

The sample supplied by the  $\beta$  network may be proportional to either load voltage or load current, as shown in Fig 5.2. In Fig 5.2a the output voltage is sampled by connecting the feedback network *in shunt* across the output. This type of connection is referred to as a *voltage, or node, sampling*. Another feedback connection which samples the output current is shown in Fig.5.2b, where the feedback network is connected *in series* with the output. This type of connection is referred to as *current, or loop, sampling*. Other sampling networks are possible.

## The Feedback Equation

The relationships between various signals in the feedback system of Fig.5.1 can be expressed as follows. The signal input to the basic amplifier,  $V_i$ , is

$$V_i = V_s + V_f \quad (5.1)$$

The output signal of the basic amplifier is

$$V_o = AV_i \quad (5.2)$$

where  $A$  is the voltage gain of the basic amplifier. The feedback signal  $V_f$  is

$$V_f = \beta V_o, \quad (5.3)$$

where  $\beta$ , the feedback factor, is the fraction of the output signal which is fed back. The terminal or closed-loop voltage gain with feedback  $A_{vf}$  is given by

$$A_{vf} = \frac{V_o}{V_s} \quad (5.4)$$

Substituting equations (5.1) and (5.3) in (5.4), we obtain

$$A_{vf} = \frac{V_o}{V_i - V_f}$$

Dividing by  $V_o$  yields

$$A_{vf} = \frac{1}{\frac{1}{A} - \beta}$$

Now we multiply by  $A$  to obtain a *generalized form of the feedback equation*. That is

$$A_{vf} = \frac{A}{1 - A\beta} \quad (5.5)$$

## Loop gain

The signal  $V_i$  in Fig. 5.1 is multiplied by  $A$  in passing through the amplifier, it is multiplied by  $\beta$  in transmission through the feedback network, and is multiplied by  $-1$  in the summing network. Such a path takes us from the input terminals around the loop consisting of the amplifier and the feedback network back to the input. The product



$$T = A\beta$$

is called the *loop gain* or *return ratio*. The difference between unity and loop gain is called the *return difference*.

$$D = 1 - A\beta$$

It should be remembered that all the quantities in equation (5.5) are functions of frequency. The  $(1 - A\beta)$  term in the denominator of equation (5.5) is a complex number. Therefore, the magnitude and phase angle of gain of the amplifier with feedback will differ from the gain of the amplifier without feedback.

### 5.3 Positive and Negative Feedback

The terms positive feedback and negative feedback are used to denote the type of feedback found in certain electronic circuits.

It, in equation (5.5),  $|1 - A\beta| < 1$ , then the overall gain  $A_{vf}$  will be greater than the gain without feedback and the feedback is said to be positive or regenerative. Of course, both the magnitude and phase angle  $\theta$  of the product  $A\beta$  are important in determining whether the feedback is positive. Generally, a regenerative circuit gives increased gain but decreased stability and higher distortion.

**Example 5.1** An amplifier has a voltage gain without feedback of  $A_v = 40$ , and 1% of the output is fed back so as to increase the input. Then  $\beta = +0.01$ , and the gain with feedback is

$$A_{vf} = \frac{40}{1 - 0.4} = 66.6$$

Since the voltage gain has been increased, the question may arise why we hardly use it. Because of reduced stability with positive feedback, this method is seldom used in an amplifier.

To illustrate the instability in an amplifier with positive feedback, we consider the following situation: No signal is applied, but due to some transient disturbance, a signal  $V_o$  appears at the output terminals. A portion of this signal,  $\beta V_o$ , will be fed back to the input terminal, and will appear in the output as an increased signal  $A\beta V_o$ . If this term just equals  $V_o$ , then the spurious output has regenerated itself. In other words, if  $A\beta V_o = V_o$  (i.e. if  $A\beta = 1$ ), then the amplifier will oscillate. Hence, if an attempt is made to obtain a large gain by making  $(A\beta)$  almost equal to unity, there is the possibility that the amplifier may break into spontaneous oscillations. This would occur if, because of variation in supply voltages, aging of transistors, etc.,  $A\beta$  becomes equal to unity, there is little point in attempting to achieve amplification at the expense of stability. In fact, feedback in amplifiers is almost always negative.

If  $|1 - A\beta| > 1$ , then the gain  $A_{vf}$  is less than  $A$ . The feedback is then said to be negative and the circuit is degenerative. A degenerative circuit, in general, increases stability, lowers gain, reduces distortion and noise in an amplifier. There are various other advantages of



negative feedback, for example, the normally high input resistance of a voltage amplifier can be made higher, and its normally low output resistance can be lowered. Another important advantage of the proper use of negative feedback is the significant improvement in the frequency response and in the linearity of operation of the feedback amplifier compared with that of the amplifier without feedback.

**Example 5.2** Determine the closed-loop gain of a feedback system having an open-loop gain of  $-10,000$  and a feedback network with  $\beta = 0.01$ .

**Solution—**  $A_{vf} = \frac{-10^4}{1 + 10^4 \times 10^{-2}} = -99$

If we use the approximation  $A\beta \gg 1$ , as in this case, then

$$A_{vf} \approx -\frac{1}{\beta} = -100$$

(note that the gain is independent of the complex gain of the internal amplifier).

### Gain & Sensitivity

An important result of negative feedback is the fact that the closed-loop gain  $A_{vf}$  is not as sensitive to parameter changes as in the open-loop gain  $A$ . The feedback network of Fig. 5.1 is regarded as a precision network, where the feedback factor  $\beta$  is accurately established and is insensitive to changes in the environment. In contrast, the gain of the basic amplifier is generally not a precise value, being highly dependent on temperature and parameter changes, such as  $h_{fe}$ .

We consider a small change in  $A_{vf}$  ( $\equiv dA_{vf}$ ) resulting from a small change in  $A$  ( $\equiv dA$ ). Then using the form for  $A_{vf}$  given in equation (5.5)

$$\begin{aligned} \frac{dA_{vf}}{dA} &= \frac{d}{dA} \left( \frac{A}{1 - A\beta} \right) \\ &= \frac{1}{(1 - A\beta)^2} \\ \frac{dA_{vf}}{A_{vf}} &= \frac{dA}{(1 - A\beta)^2} \left( \frac{1 - A\beta}{A} \right) \\ &= \frac{dA}{A} \left( \frac{1}{1 - A\beta} \right) \end{aligned} \quad (5.6)$$

Thus fractional change in the closed-loop gain ( $dA_{vf}/A_{vf}$ ) resulting from the sensitivity of the basic amplifier has been reduced through the application of negative feedback by the factor



$$D^1 = \frac{1}{1 - A\beta}$$

This factor is often referred to as the *sensitivity factor* of a feedback system.

**Example 5.3** In example 5.2 compute the fractional change in the overall gain for a 50 percent increase in the gain of the basic amplifier

**Solution:** From equation (5.6) with  $A = -10,000$ ,  $\beta = 0.01$

$$\begin{aligned} \frac{dA_{vf}}{A_{vf}} &= \frac{dA}{A} \left( \frac{1}{1 - A\beta} \right) \\ &= 0.5 \left( \frac{1}{1 + 100} \right) \\ &= 0.00495 \end{aligned}$$

Thus the open-loop gain change of 50% results in a closed loop gain change of less than 0.5%.

This example illustrates the gain stability of an amplifier with negative feedback.

**Example 5.4** If an amplifier with gain of -1000 and feedback of  $\beta = 0.1$  has a gain change of 20% due to temperature, calculate the change in gain of the feedback amplifier.

**Solution:** Here  $|A\beta| \gg 1$

$$\therefore \frac{dA_{vf}}{A_{vf}} = \frac{dA}{A} \frac{1}{(-\beta A)} = (20\%) \frac{1}{0.1 \times (+1000)} = 0.2\%$$

The improvement in stability is thus 100 times.

## 5.4 Barkhausen Criterion

From equation (5.5) it follows that for  $A\beta = 1$ , then  $A_{vf} \rightarrow \infty$ . This may be interpreted to mean that there exists an output voltage even in the absence of an oscillator. In general, both  $A$  and  $\beta$  are complex quantities, and  $A\beta = |A\beta| e^{i\theta}$  should be used. Here  $\theta$  is the total phase shift associated with the amplifier and the feedback loop. Thus two basic conditions for sustained oscillations are

1. The loop gain  $|A\beta| = 1$
2.  $\theta = 2n\pi$ . In other words the net phase shift around the feedback loop must be an integral multiple of  $2\pi$ .

These requirements for sustained oscillation are called Barkhausen criterion.

In all practical feedback oscillators  $|A\beta|$  must be greater than unity. Because the



condition  $|\beta A| = 1$  does not give a range of acceptable values of  $|\beta A|$ , but rather a single and precise value. Now suppose that initially it were even possible to satisfy this condition. Then, because circuit components and, more importantly, transistors change characteristics with age, temperature, voltage, replacement, etc, it is clear that if the entire oscillator is left to itself, in a very short time  $|\beta A|$  will become either less or more than unity. In the former case the oscillation simply stops, and in the latter case we are back to the point of requiring nonlinearity to limit the amplitude. An oscillator in which the loop gain is exactly unity is an abstraction completely unrealizable in practice. It is accordingly necessary, in the adjustment of a practical oscillator, always to arrange to have  $|\beta A|$  somewhat larger (say 5 percent) than unity in order to ensure that, with incidental variations in transistor and circuit parameters,  $|\beta A|$  shall not fall below unity. While the first two principles stated above must be satisfied on purely theoretical grounds, we may add a third general principle dictated by practical considerations, i.e.

3. In every practical oscillator the loop gain is slightly larger than unity, and the amplitude of the oscillations is limited by the onset of nonlinearity.

## 5.5 Oscillators

The electronic device employed for generation of electrical oscillations is called an *electronic oscillator* and the circuit is called oscillator circuit. In this device dc power is converted into ac power and provides a constantly varying output signal. If the output signal varies sinusoidally, the circuit is referred to as a *sinusoidal oscillator*. There is another class of oscillators in which the output is non-sinusoidal and is commonly known as *relaxation oscillators*. The types most commonly used are: (i) multivibrator, (ii) square wave generator, (iii) pulse generator, (iv) saw-tooth generator. These non-sinusoidal oscillators are generally used as electronic timing and control circuits in television, radar, cathode ray oscilloscope, and control equipments.

To achieve sustained oscillations positive feedback is used and Barkhausen criterion must be satisfied. The essential components of an oscillator are shown in Fig. 5.3 :

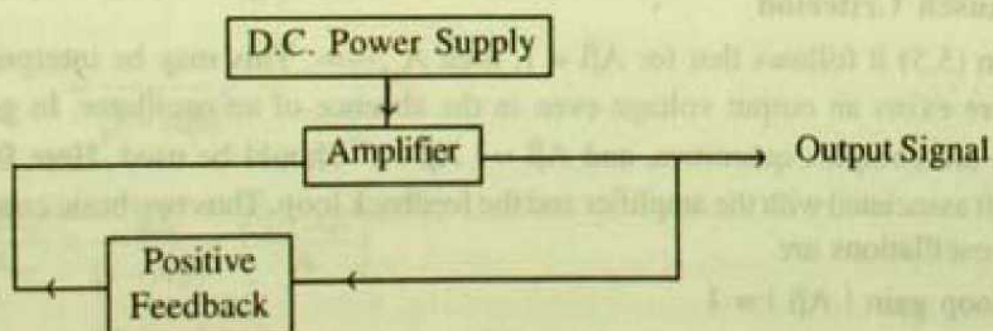


Fig. 5.3 Block diagram of a feedback oscillator.



## 5.6 Questions and problems

1. Draw a feedback amplifier in block diagram form. Identify each block, and state its function.
2. Define (a) feedback factor, (b) negative feedback, (c) positive feedback, (d) desensitivity factor.
3. What is Barkhausen criterion ?
4. For a feedback system having a closed-loop gain of 100 and an open-loop gain of 2000, determine the desensitivity factor and compute the fractional change in the close-loop gain resulting from a 20% change in the open-loop gain.

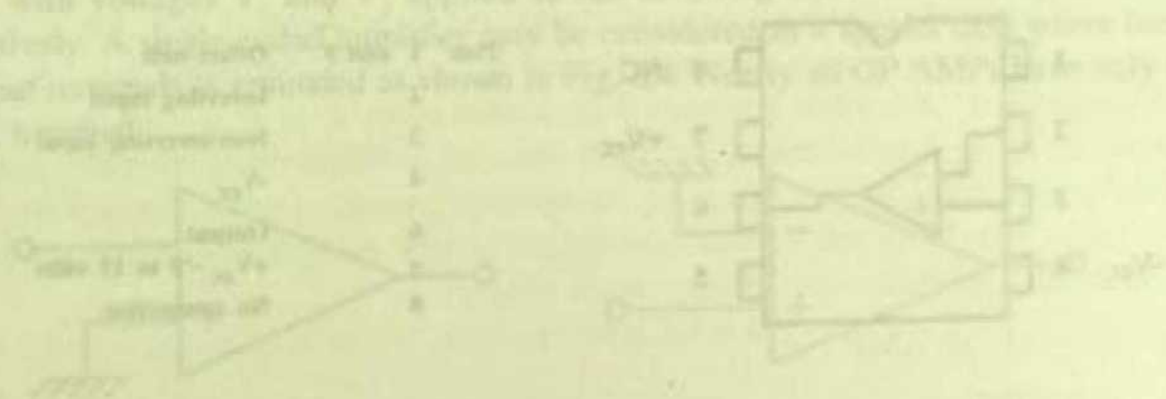


Fig. 5.1 The connecting of OP AMP (BET. 741). The triangle within the box is the symbol of OP AMP. (DAA-20) (BET. 741)

## Chapter 6

### Operational Amplifiers

#### 6.1 Introduction

The term Operational Amplifier (abbreviated OP AMP) originally referred to a class of amplifiers that could be used in various feedback configurations to perform a number of mathematical operations such as addition, multiplication, differentiation, integration etc. Although OP AMPs were originally developed to perform these mathematical operations, their versatility has made them useful in a host of other applications.

OP AMPs are used, together with external components (passive or active), in such a way that the performance of the complete unit is primarily a function of the external components and we need not bother about the internal complicated electronic circuitry of the OP AMPs. Thanks to the integrated circuit (IC) technology, which has revolutionized the modern civilization. IC technology has reduced the size of a complete OP AMP by a factor which is hard to believe. The size of IC OP AMPs for normal use is governed by ease of handling and the need to make electrical connections to the unit, rather than the actual size of the IC itself ( $\sim 1\text{mm}^2$ ). It may be mentioned that BEL (Bharat Electronics Limited) IC 741 OP AMP has 11 npn, 6 pnp transistors, 10 resistors, 4 diodes and 1 capacitor. This IC OP AMP is marketed with eight pin connections in the form of either the hermetically sealed metal can with a circular array of wire leads or a cheaper one dual-in-line (DIL or DIP) package with two parallel rows of pins (Fig 6.1). The metal can has the advantage of somewhat higher power dissipation, while the DIP has a moulded plastic body.

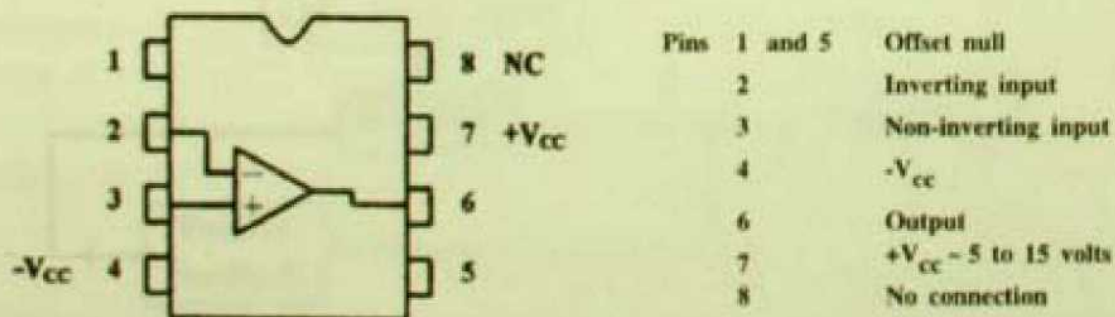


Fig. 6.1 Pin connecting of OP AMP (BEL IC 741). The triangle within the box is the symbol of OP AMP



The schematic diagram of an OP AMP is shown in Fig 6.2 and the equivalent circuit in Fig.6.3. The meaning of the minus(-) and plus(+) signs appearing at the input of the OP AMP is that the (-) sign indicates the *inverting input* terminal, i.e. the difference

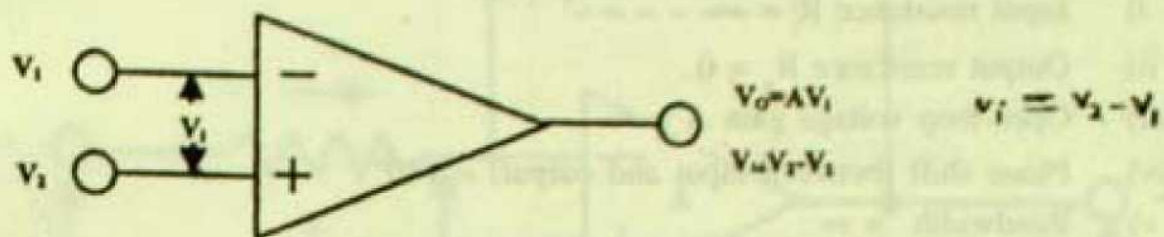


Fig. 6.2

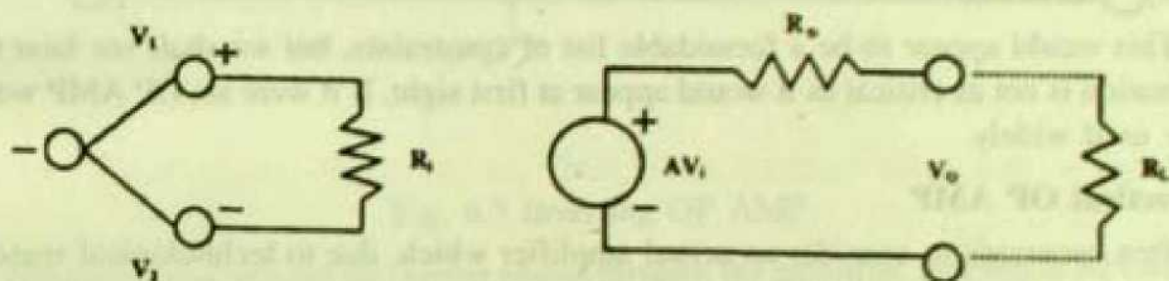


Fig. 6.3 Equivalent circuit of OP AMP

between input and output is  $180^\circ$  while the (+) sign indicates the *noninverting input* terminal i.e. the phase difference between input and output is  $0^\circ$  or  $360^\circ$ . In other words, a *positive-going* signal  $V_i$  appearing at the (-) input (with reference to the (+) input) would cause the output of the voltage source  $V_o$  to have the amplified *negative-going* signal  $-AV_i$ . Referenced to the (-) input, a positive-going signal at the (+) input would cause a positive going signal at the output of the amplifier. A large number of OP AMPs have a differential input, with voltages  $V_1$  and  $V_2$  applied to the inverting and noninverting terminals, respectively. A single ended amplifier may be considered as a special case where one of the input terminals is grounded as shown in Fig. 6.4. Nearly all OP AMPs have only one output terminal.

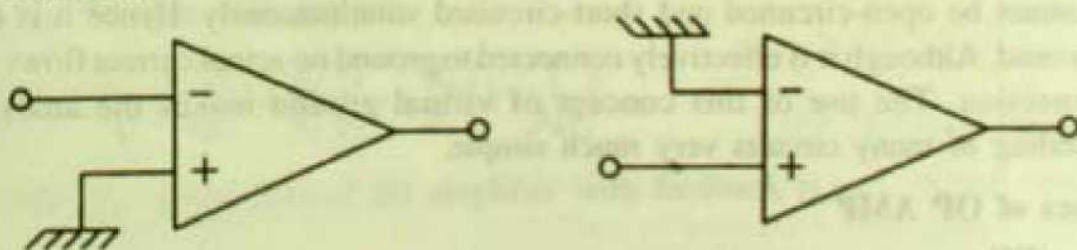


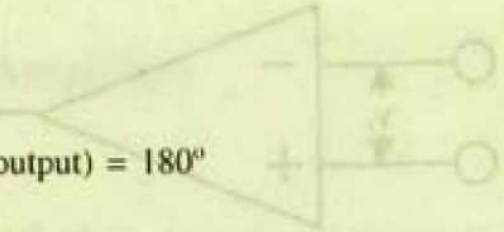
Fig. 6.4 Single-ended OP-AMP

## 6.2 Characteristics

### A. Ideal OP APM

To demonstrate the basic points of OP AMPs we shall first consider the ideal amplifier. This will have the following characteristics.

- i) Input resistance  $R_i = \infty$
- ii) Output resistance  $R_o = 0$
- iii) Open-loop voltage gain  $A = \infty$
- iv) Phase shift (between input and output) =  $180^\circ$
- v) Bandwidth =  $\infty$
- vi) Perfect balance  $V_o = 0$  for  $V_1 = V_2$
- vii) Zero off-set
- viii) Zero noise
- ix) Zero drift



This would appear to be a formidable list of constraints, but we shall see later that the situation is not as critical as it would appear at first sight. If it were so, OP AMP would not be used widely.

### B. Practical OP AMP

It is often necessary to consider an actual amplifier which, due to technological reasons, cannot satisfy the ideal conditions, viz.  $|A| \neq \infty$ ,  $R_i \neq \infty$ ,  $R_o \neq 0$ . In the case of 741 IC chip, input impedance  $R_i = 1-3$  megohm, output impedance  $R_o = 75-100$  ohm. Gain bandwidth product  $f_T = 10^6$ , common mode rejection ratio = 80 dB

## 6.3 Concept of virtual ground

The vital concept arising out of the idealization is that of the *virtual ground*. By definition,  $V_i = V_o / A = 0$ , since  $V_o$  is finite (maximum value is  $V_{cc}$  – the supply voltage), and  $A = \infty$ ; hence  $V_i = 0$ . Thus we may say that the input terminal of the amplifier is maintained at ground of potential, i.e. *short circuited*. Again, the input current  $I_i = V_i / R_i = 0$ , which means that the input is *open-circuited*. To sum up, we see that the input is at ground from voltage point of view, while it is open from current point of view. A point in an electrical circuit cannot be open-circuited and short-circuited simultaneously. Hence it is called a virtual ground. Although it is effectively connected to ground no actual current flows through this connection. The use of this concept of virtual ground makes the analysis and understanding of many circuits very much simple.

## 6.4 Uses of OP AMP

### a. Amplifiers

#### a.1 Inverting OP AMP

The circuit diagram of an inverting OP AMP is shown (Fig. 6.5).



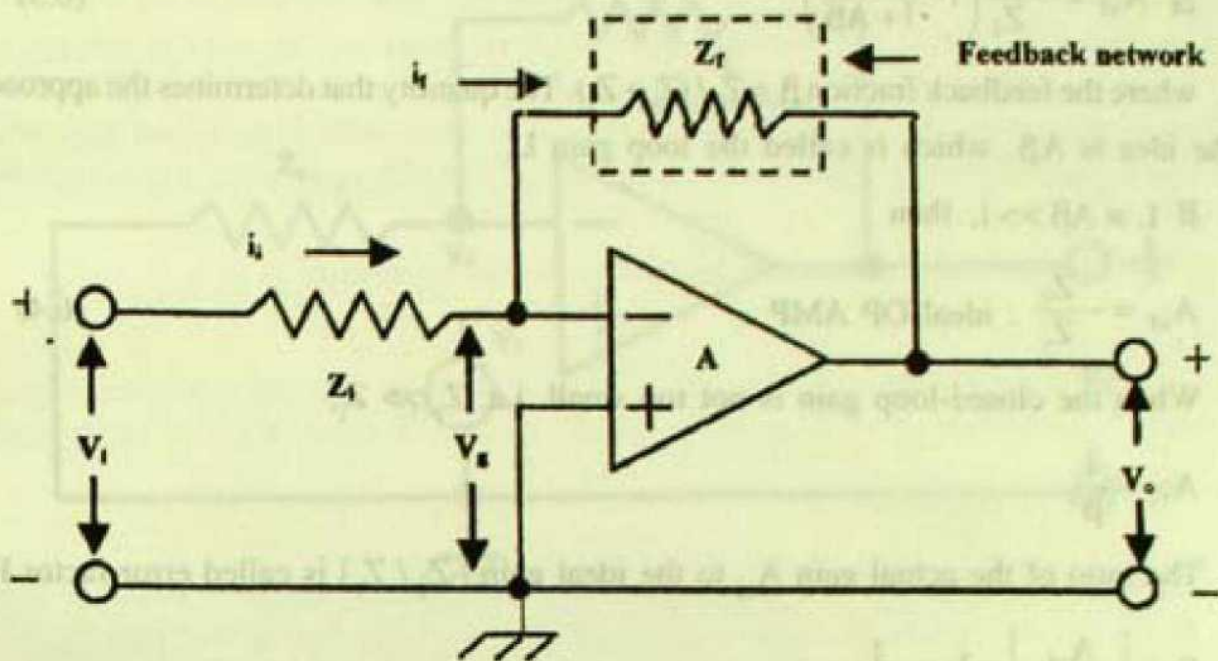


Fig. 6.5 Inverting OP AMP

Due to virtual ground no current passes through the amplifier, and hence we can write

$$i_i = i_f \quad (6.1)$$

$$V_o = -AV_i \quad (6.2)$$

Now using Kirchhoff's Voltage Law (KVL) we substitute for  $i_i$  and  $i_f$  in equation (6.1) as

$$\frac{V_i - V_g}{Z_1} = \frac{V_g - V_o}{Z_f}$$

$$\text{or } \frac{V_i}{V_o} - \frac{V_g}{V_o} = \frac{Z_1}{Z_f} \left( \frac{V_g}{V_o} - 1 \right)$$

Using equation (6.2) we get

$$\frac{V_i}{V_o} = -\frac{Z_1}{Z_f} \left[ 1 + \frac{1}{A} \left( 1 + \frac{Z_f}{Z_1} \right) \right] = -\frac{Z_1}{Z_f} \left( 1 + \frac{1}{A\beta} \right)$$

Hence, the actual gain of the amplifier with feedback is

$$A_{vt} = \frac{V_o}{V_i} = -\frac{Z_f}{Z_1} \frac{1}{1 + \frac{1}{A\beta}}$$

$$\text{or } A_{vf} = -\frac{Z_f}{Z_i} \left( 1 - \frac{1}{1 + A\beta} \right) \quad (6.3)$$

where the feedback fraction  $\beta = Z_i / (Z_i + Z_f)$ . The quantity that determines the approach to the idea is  $A\beta$ , which is called the loop gain  $L$ .

If  $L \equiv A\beta \gg 1$ , then

$$A_{vf} = -\frac{Z_f}{Z_i} : \text{ideal OP AMP} \quad (6.4)$$

When the closed-loop gain is not too small, i.e.  $Z_f \gg Z_i$ ,

$$A_{vf} \approx \frac{1}{\beta}.$$

The ratio of the actual gain  $A_{vf}$  to the ideal gain  $|Z_f / Z_i|$  is called error factor  $F_e$

$$F_e = \left| \frac{A_{vf}}{Z_f / Z_i} \right| = 1 - \frac{1}{1 + A\beta}.$$

To get an idea about feedback in OP AMP we may rewrite equation (6.3) as

$$\begin{aligned} A_{vf} &= -\frac{AZ_f}{Z_f + (1 + A)Z_i} \\ &= -\frac{A}{1 - \left( -A \frac{Z_i}{Z_f} \right)}, \end{aligned} \quad (6.5)$$

since, in general,  $A \gg 1$ . In terms of feedback equation,

$$A_{vf} = \frac{a}{1 - af}$$

From equation (6.5)

$$a = -A$$

$$f = \frac{Z_i}{Z_f}.$$

## a.2 Noninverting OP AMP

As the name implies, in a noninverting amplifier the output signal is of the same polarity as the input signal, only increased in magnitude; there is no polarity inversion. In this configuration the input signal is applied to the positive terminal of the amplifier (as shown in Fig. 6.6). The feedback is returned to the negative terminal so that the feedback is in proper polarity. Then the feedback voltage  $V_f$  will be a fraction of the output voltage  $V_o$ .



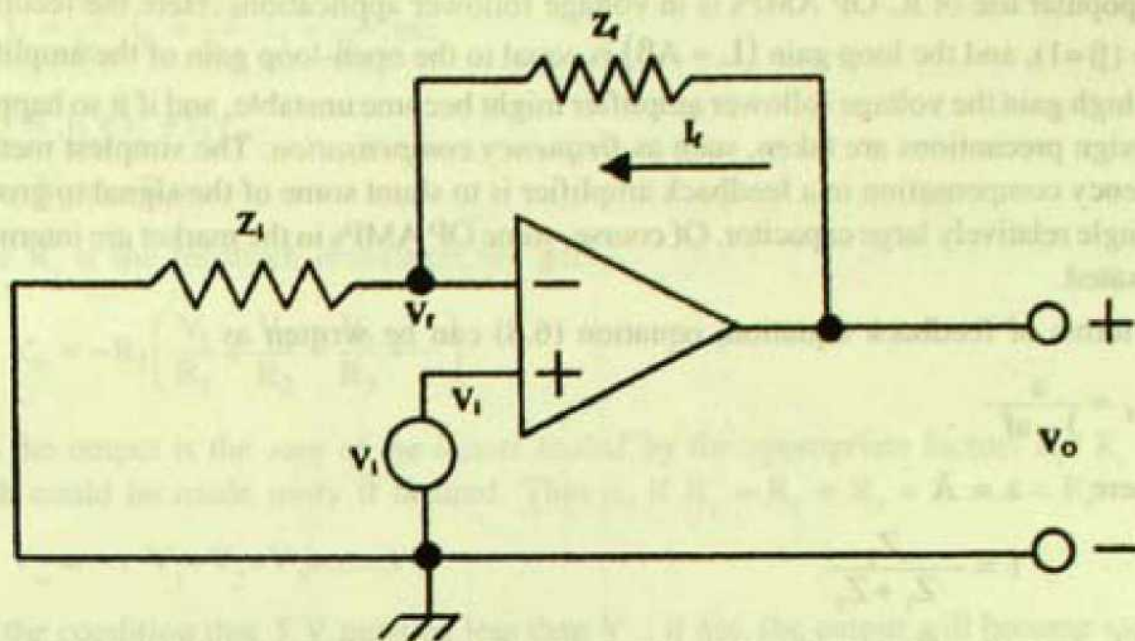


Fig. 6.6 Non-inverting OP AMP

$$V_f = \frac{Z_i}{Z_i + Z_f} V_o \quad (6.6)$$

The differential input voltage to the amplifier is  $V_i - V_f$  and the output of the amplifier is

$$\begin{aligned} V_o &= A(V_i - V_f) \\ &= A \left( V_i - \frac{Z_i}{Z_i + Z_f} V_o \right), \end{aligned} \quad (6.7)$$

substituting the value of  $V_f$  from equation. (6.6). The gain can now be written as, from equation. (6.7)

$$A_{vf} = \frac{V_o}{V_i} = \frac{A}{1 + AZ_i / (Z_f + Z_i)} \quad (6.8)$$

$$\text{or } A_{vf} = \frac{1}{\frac{Z_i}{Z_i + Z_f} + \frac{1}{A}} \quad (6.9a)$$

For ideal OP AMP,  $A \rightarrow \infty$ . In this case

$$A_{vf} = 1 + \frac{Z_f}{Z_i} \quad (6.9b)$$

Hence, the closed-loop gain is always greater than unity. If  $Z_i = \infty$  and/or  $Z_f = 0$ , then  $A_{vf} = 1$  ( $\equiv 0$  dB) and the amplifier acts as a *Voltage follower*.

A very popular use of IC OP AMPs is in voltage follower applications. Here the feedback is 100% ( $\beta=1$ ), and the loop gain ( $L = A\beta$ ) is equal to the open-loop gain of the amplifier. At such high gain the voltage follower amplifier might become unstable, and if it so happens other design precautions are taken, such as *frequency compensation*. The simplest method of frequency compensation in a feedback amplifier is to shunt some of the signal to ground with a single relatively large capacitor. Of course, some OP AMPs in the market are internally compensated.

In terms of feedback equation, equation (6.8) can be written as

$$A_{vf} = \frac{a}{1 - af}$$

where  $a = A$

$$f = -\frac{Z_f}{Z_i + Z_f}$$

Aside from its effect on the gain properties of an amplifier, negative feedback also has an effect on the input and output resistances. For the treatment of this part student may go through any advanced textbook on electronics.

## b. Mathematical operations

### b.1 The adding operation or Adder

The existence of virtual ground makes it possible to apply several inputs simultaneously without interaction. The output will then be the algebraic sum of the inputs multiplied by some constants which may be chosen to be unity. The circuit for the adder is shown in Fig. 6.7.

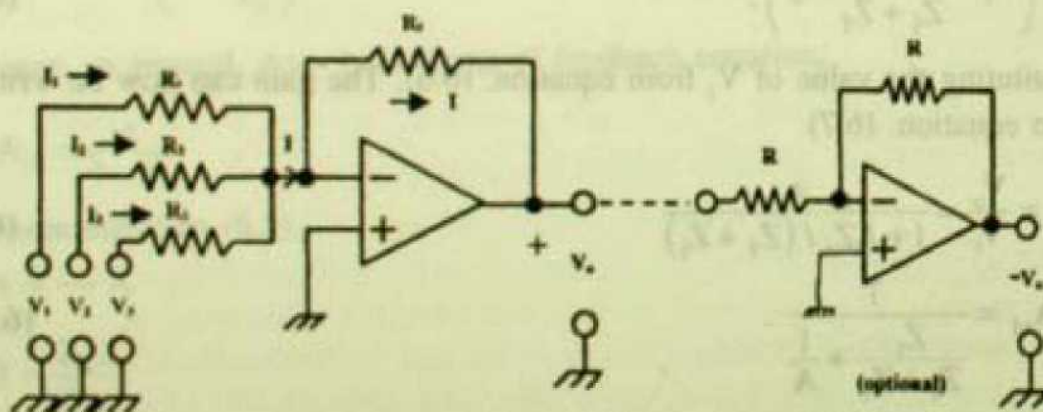


Fig. 6.7 Adding OP AMP

The adding point at the amplifier inverting input terminal is virtually at ground potential, so the currents  $i_1, i_2, i_3, \dots$  will follow through the respective input resistors  $R_1, R_2, R_3, \dots$  uninfluenced by each other and driven by voltage sources  $V_1, V_2, V_3, \dots$ . From the figure 6.7 we see that



$$i_1 = \frac{V_1}{R_1}, i_2 = \frac{V_2}{R_2}, i_3 = \frac{V_3}{R_3}, \text{etc.}$$

$$\text{and } i = i_1 + i_2 + i_3 + \dots$$

$$\text{Since } V_o = -R_f i$$

where  $R_f$  is the feedback resistance, we get

$$V_o = -R_f \left( \frac{V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_3}{R_3} + \dots \right) \quad (6.10)$$

Thus the output is the *sum of the inputs scaled* by the appropriate factors  $R_f / R_1, \dots$  etc. which could be made unity if desired. That is, if  $R_1 = R_2 = R_3 = \dots = R_f$

$$V_o = - (V_1 + V_2 + V_3 + \dots) \quad (6.11)$$

with the condition that  $\sum V_i$  must be less than  $V_{cc}$ ; if not, the output will become saturated. Thus the circuit is that of an *adding* circuit with a change in sign which may be inverted by another inverting OP AMP with unity gain.

Many other methods may, however, be used to combine signals or voltages. The present method has the advantage that it may be extended to any number of inputs requiring only one additional resistor for each additional input. The result depends, in the limiting case of large amplifier gain, only on the resistors involved and due to the virtual ground, there will be a minimum of interaction between input sources.

By making the ratio  $R_f / R = 1/n$ , we can obtain an average of  $n$  signal voltages.

For addition of several ac signals the adder amplifier should have a flat frequency response. To obtain sufficient bandwidth, some amplifiers may have to sacrifice gain because the gain bandwidth product of an amplifier is constant.

**Example 6.1** What is the output voltage of an OP AMP adder for the following sets of input voltages and resistors?  $R_f = 100 \text{ k}\Omega$  in all cases.

$$V_1 = -2\text{V}, \quad V_2 = +3\text{V}, \quad V_3 = +1.5\text{V}$$

$$R_1 = 10 \text{ k}\Omega, \quad R_2 = 50 \text{ k}\Omega, \quad R_3 = 10 \text{ k}\Omega$$

**Solution,**

$$V_o = - \left[ \frac{100}{10} (-2\text{V}) + \frac{100}{50} (+3\text{V}) + \frac{100}{10} (1.5\text{V}) \right]$$

$$= +1\text{V}$$

## b.2 OP AMP Subtractor

The general case with unrelated resistors gives rather unwieldy expression, and hence we shall consider the case shown in Fig. 6.8. In the figure,  $V_+$  and  $V_-$  are the voltages at

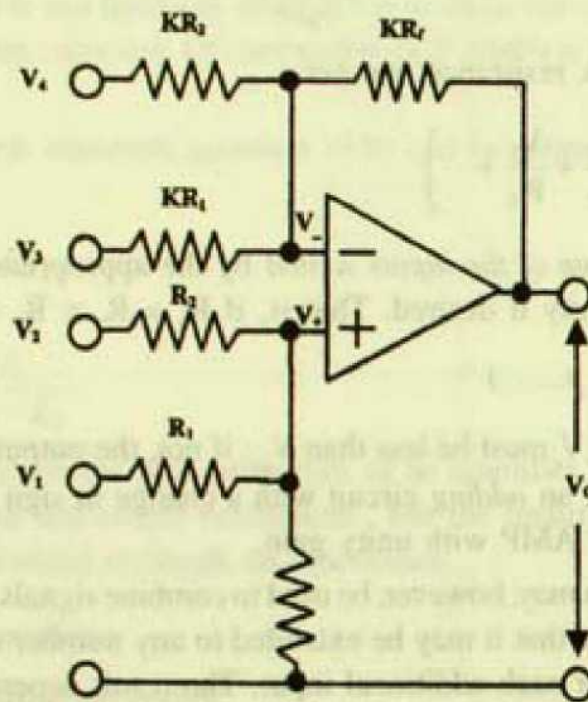


Fig. 6.8 OP AMP Subtractor

the inverting and noninverting input respectively, and  $K$  is a numerical factor. For the (-) input :

$$\frac{V_4 - V_-}{KR_2} + \frac{V_3 - V_-}{KR_1} = \frac{V_- - V_o}{KR_f}$$

$$\text{or, } V_- \left[ \frac{1}{R_f} + \frac{1}{R_1} + \frac{1}{R_2} \right] = \frac{V_4}{R_2} + \frac{V_3}{R_1} + \frac{V_o}{R_f} \dots \quad (6.12)$$

For the (+) input :

$$\frac{V_2 - V_+}{R_2} + \frac{V_1 - V_+}{R_1} = \frac{V_+}{R_f}$$

$$\text{or, } V_+ \left[ \frac{1}{R_f} + \frac{1}{R_1} + \frac{1}{R_2} \right] = \frac{V_1}{R_1} + \frac{V_2}{R_2} \quad (6.13)$$

Since  $V_+ - V_- = V_g = 0$  i.e.  $V_+ = V_-$ . Hence, from equation (6.12) and (6.13)

$$V_o = R_f \left[ \frac{V_1}{R_1} + \frac{V_2}{R_2} - \frac{V_3}{R_1} - \frac{V_4}{R_2} \right] \quad (6.14)$$



The above configuration requires that there must be an equal number of (+) and (-) inputs; this can always be done, the unused inputs being grounded, so that  $V_n = 0$ .

### b.3 Analog integration or OP AMP Integrator

By use of a capacitor  $C$  in place of  $Z_f$ , and a resistor  $R$  in place of  $Z_i$ , the OP AMP in the inverting mode will perform the mathematical operation of *integration* on the input signal (Fig. 6.9)

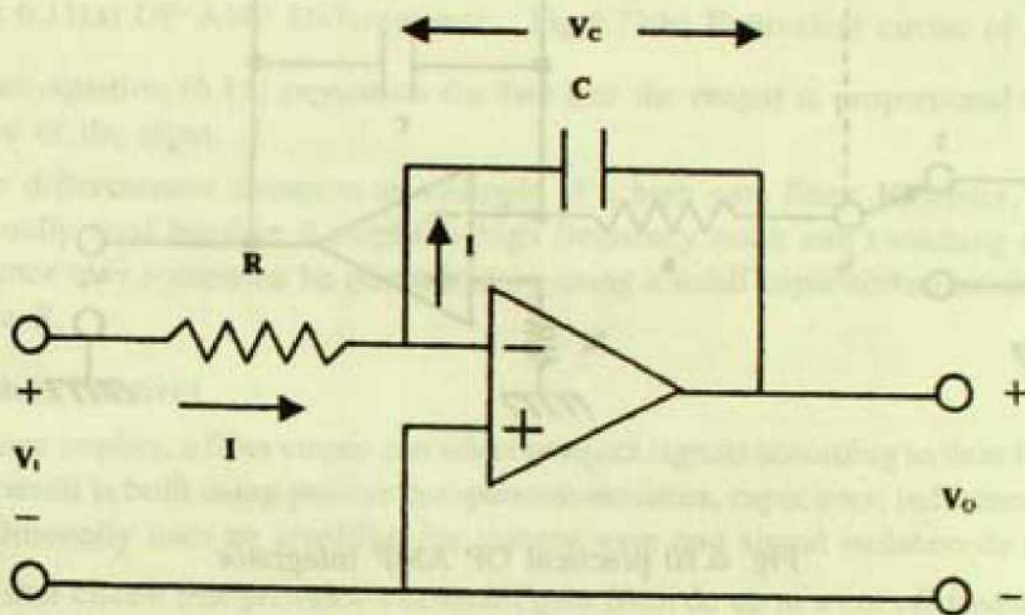


Fig. 6.9 OP AMP Integrator

We have the output voltage  $V_o = -V_c$ , where the capacitor voltage  $V_c$  is related to its charge  $q$  by

$$V_c = \frac{q(t)}{C}.$$

Since the capacitor charge is the integral of the current  $i(t)$ , we may write

$$q(t) = \int_0^t i(t_1) dt_1. \text{ Hence}$$

$$V_o(t) = -V_c(t) = -\frac{1}{C} \int_0^t i(t_1) dt_1$$

$$\text{or, } V_o(t) = -\frac{1}{RC} \int_0^t V_i(t_1) dt_1 \quad (6.14)$$

The output voltage is thus the integral of the input voltage multiplied by a negative constant. The negative sign can be eliminated by another stage of amplifier (viz. inverting amplifier with gain 1). If we make  $R = 1\text{M}\Omega$  and  $C = 1\mu\text{F}$ , the constant becomes unity.

A practical integrator, however, must be provided with an external circuit to introduce the initial conditions, as shown in Fig. 6.10. Switches  $S_a$  and  $S_b$  are coupled which means

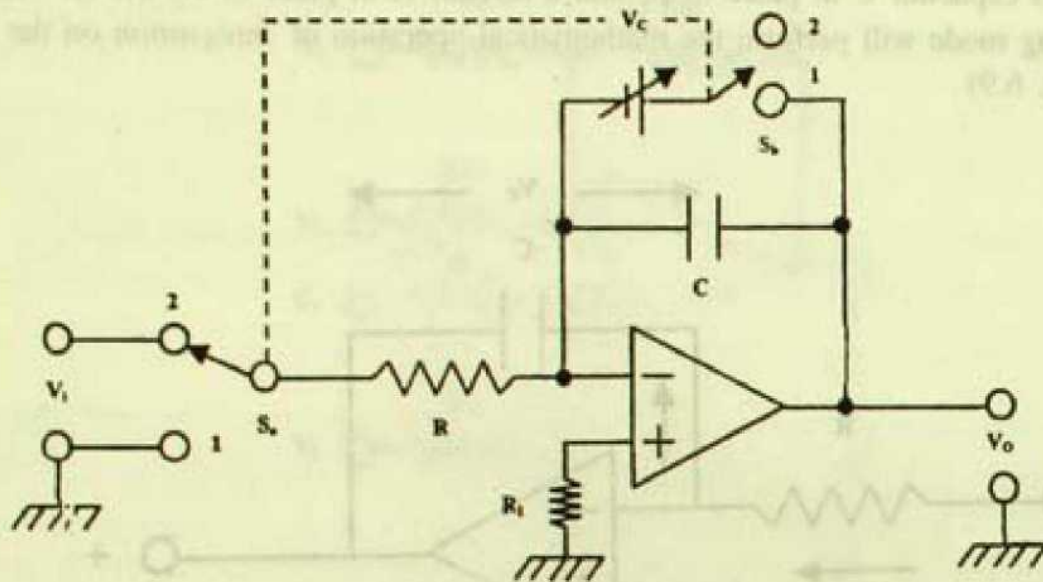


Fig. 6.10 practical OP AMP Integrator

they have double poles and double throws and hence such switches are called DPDT. When the switch  $S_a$  is in position 1, the input is zero and the capacitor  $C$  is charged to the voltage  $V_c$ , setting an initial condition for  $V_o$ . When the switch is in position 2, the amplifier is connected as an integrator and its output will be  $V_c$  plus a constant (i.e.  $1/RC$ ) times the integral of the input voltage  $V_i$ . In using this circuit, care must be taken to stabilize the amplifier and  $R_i$  must be equal to  $R$  to minimize the error due to bias current.

It may be remarked that the integrator circuit is a simple example of a *low pass filter*. These circuits are extensively used in digital voltmeters and in general for the generation of voltage ramps (i.e. voltage  $\propto$  time) used as time bases in CRO.

#### b.4 OP AMP Differentiator

If the positions of  $R$  and  $C$  are interchanged in an OP AMP integrator, the resulting circuit becomes a *differentiator* (as shown in Fig. 6.9). We see from the equivalent circuit of Fig. 6.11 (a) that

$$i = \frac{dq}{dt} = C \frac{dV_i}{dt}$$

and  $V_o = -Ri = -RC \frac{dV_i}{dt}$  (6.15)



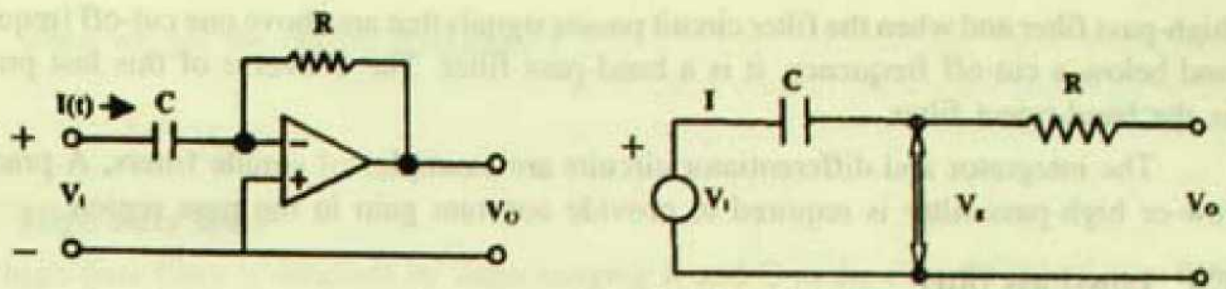


Fig.6.11(a) OP AMP Differentiator Fig.6.11(b) Equivalent circuit of Fig.6.11(a)

Thus equation (6.15) represents the fact that the output is proportional to the time derivative of the input.

The differentiator circuit is an example of a high pass filter. However, this circuit is not usually used because it amplifies high frequency noise and switching spikes. The performance may somewhat be minimized by using a small capacitor across the feedback resistance  $R$ .

### c. Filters (Active)

As the name implies, a filter circuit can select or reject signals according to their frequencies. A filter circuit is built using passive components-resistors, capacitors, inductors. An active filter additionally uses an amplifier for voltage gain and signal isolation or buffering.

A filter circuit that provides a constant gain from dc up to a cut off frequency  $f_c$  and then passes no signal is an ideal low-pass filter, the frequency response curve of which is shown is Fig. 6.16(a). A filter that passes signals only above a cut off frequency is a

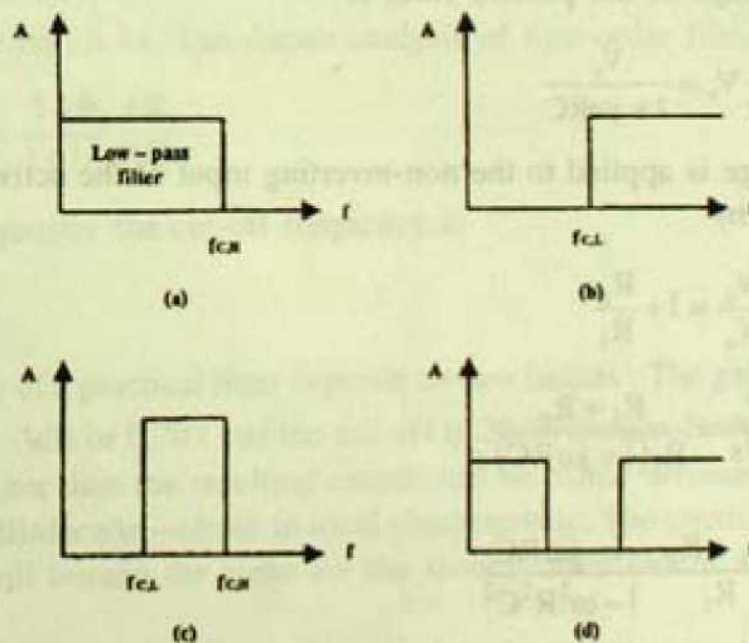


Fig. 6.12 Ideal filter response : (a) low-pass, (b) high-pass, (c) band-pass, (d) band-reject

high-pass filter and when the filter circuit passes signals that are above one cut-off frequency and below a cut-off frequency, it is a band-pass filter. The converse of this last process is the band-reject filter.

The integrator and differentiator circuits are examples of simple filters. A practical low-or high-pass filter is required to provide constant gain in the pass region.

### C.1 Low-Pass filter

A first-order low-pass filter using a single resistor and capacitor is shown in Fig. 6.13.

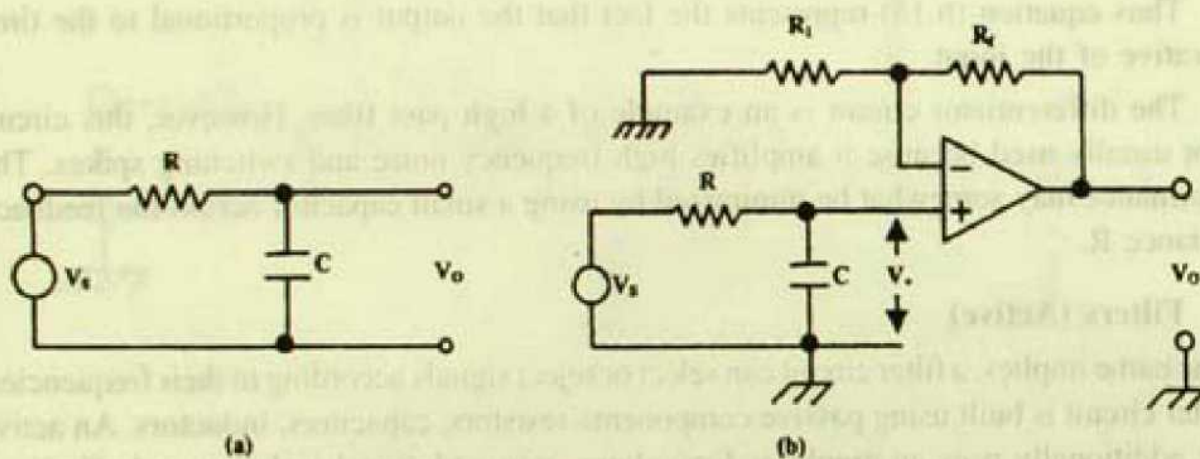


Fig. 6.13 Low-pass filter : (a) passive, (b) active

The output voltage of the passive filter is

$$V_+ = \frac{1/j\omega C}{R + 1/j\omega C} V_s = \frac{V_s}{1 + j\omega RC} \quad (6.18)$$

Now this voltage is applied to the non-inverting input of the active filter. The gain is (see equation 6.9 b)

$$A'_{vf} = \frac{V_o}{V_+} = 1 + \frac{R_f}{R_i}$$

and

$$A_{vf} = \frac{V_o}{V_s} = \frac{R_i + R_f}{R_i(1 + j\omega RC)}$$

$$= \frac{R_i + R_f}{R_i} \cdot \frac{1 - j\omega RC}{1 - \omega^2 R^2 C^2} \quad (6.19)$$

This gain will be maximum when the imaginary term is zero, i.e.

$$\omega_c RC = 0$$



when we get the cut-off frequency  $f_c$  :

$$f_c = \frac{1}{2\pi RC} \quad (6.20)$$

## C.2 High-Pass filter

The high-pass filter is obtained by interchanging R and C in the circuit for low pass filter

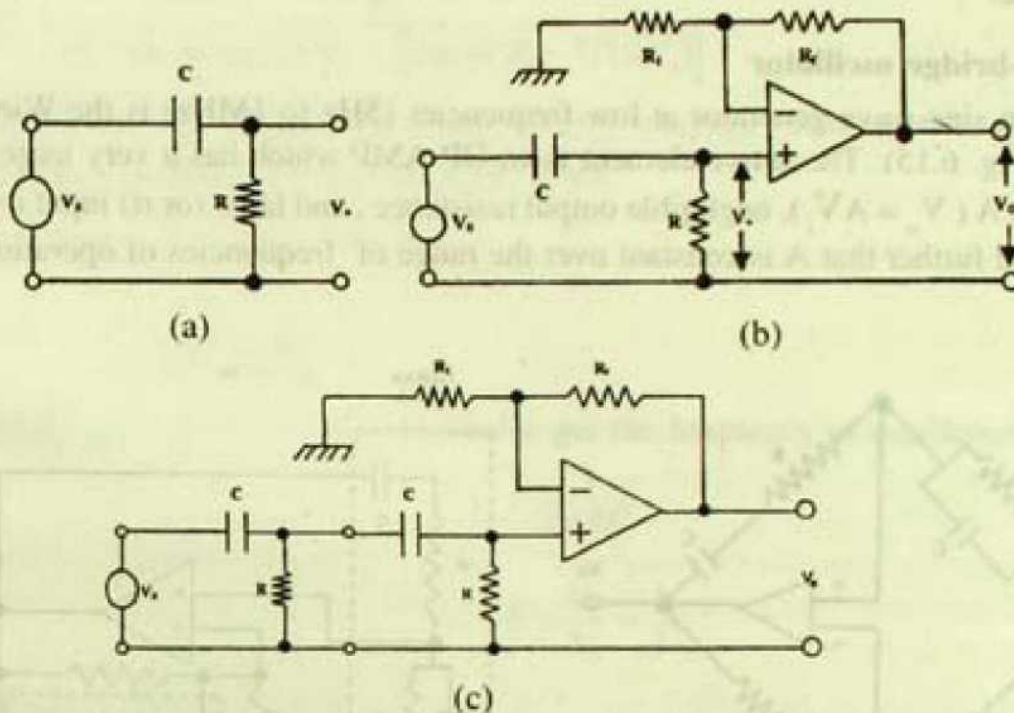


Fig. 6.14 High-pass filter : (a) passive, (b) first-order, (c) Second Order and is shown in Fig. 6.14. The circuit analysis of first-order filter circuit gives

$$A_{vf} = \frac{V_o}{V_s} = \frac{1 + R_f / R_i}{1 + 1/jRC\omega}$$

and consequently the cut-off frequency is

$$f_c = \frac{1}{2\pi RC}$$

The quality of a practical filter depends on two factors : The gain at  $f_c$  and the roll off. The gain at  $f_c$  is -3dB or 0.707 and the roll off is 20dB/decade. If we connect two sections of the passive filter then the resulting circuit will be called *second-order* and the roll off will become 40dB/decade—closer to ideal characteristic. The circuit voltage gain and cut-off frequency will remain the same for the second-order circuit as for the first order.

## d. Oscillator

OP AMPS are conveniently used in oscillators capable of generating a variety of output waveforms. Basically, the function of an oscillator is to generate alternating current or voltage

waveform. More precisely, an oscillator is a circuit that generates a repetitive waveform of fixed amplitude and frequency without any external source. Oscillators are essential in all electronics laboratories and are used in radio, television, computers, and communication. The type of waveform generated by an oscillator depends on the components in the circuit and are mainly of four types viz., sinusoidal, square, triangular and saw-tooth. All oscillators work on the same basic principle of positive feedback. We shall consider one example of each type.

### d.1 Wien-bridge oscillator

An excellent sine-wave generator at low frequencies (5Hz to 1MHz) is the Wien bridge oscillator (Fig. 6.15). The active element is an OP AMP which has a very large positive voltage gain  $A$  ( $V_o = AV_i$ ), negligible output resistance, and large (or  $\infty$ ) input resistance. It is assumed further that  $A$  is constant over the range of frequencies of operation of this circuit.

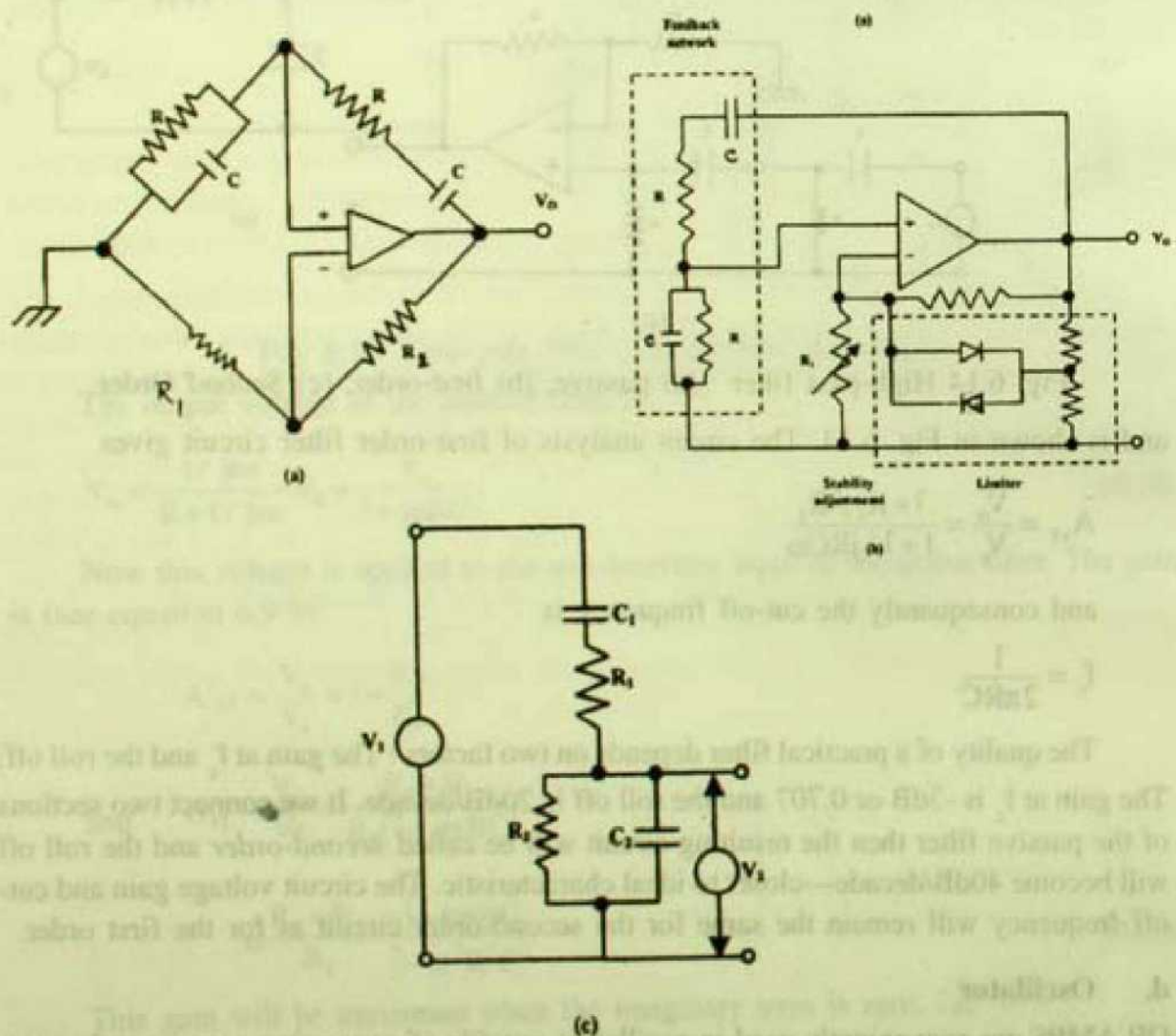


Fig. 6.15 Wien bridge oscillator (a) Principle, (b) actual circuit, (c) Wienbridge network



The feedback RC-network consists of two arms of a generalized Wheatstone bridge as shown in Fig. 6.15. One arm comprises a resistor for  $R_1$  and a capacitor  $C_1$  in series while the second arm consists of a parallel combination of  $R_2$  and  $C_2$  (Fig. 6.15c). The feedback voltage  $V_2$  is developed across the parallel arm and  $V_1$  is across the Wien-bridge network. From network analysis we find

$$\frac{V_2}{V_1} = \frac{R_2}{R_1 + R_2 + (R_2 C_2)/C_1 + j[\omega C_2 R_1 R_2 - 1/(\omega C_1)]}$$

For the network to have zero phase shift the imaginary term must be zero. Suppose, this happens at a frequency  $f_0 = \omega_0 / 2\pi$ . Then

$$\omega_0 C_2 R_1 R_2 = 1/(\omega_0 C_1)$$

$$\text{or } f_0 = \frac{1}{2\pi\sqrt{R_1 C_1 R_2 C_2}}$$

For  $R_1 = R_2$  and  $C_1 = C_2$ , we finally get the frequency of oscillation

$$f_0 = \frac{1}{2\pi RC}$$

and the feedback factor

$$\beta = \frac{V_2}{V_1} = \frac{1}{3}.$$

## d.2 Square wave generator

In contrast to sine wave oscillators square wave outputs are generated when the OP AMP is forced to operate in the saturated region. In other words, the output of the OP AMP is forced to swing repetitively between  $\pm V_{cc}$ , resulting in the square wave output. Such square wave generators are also called a *free-running* or *astable* multivibrator. Fig. 6.16

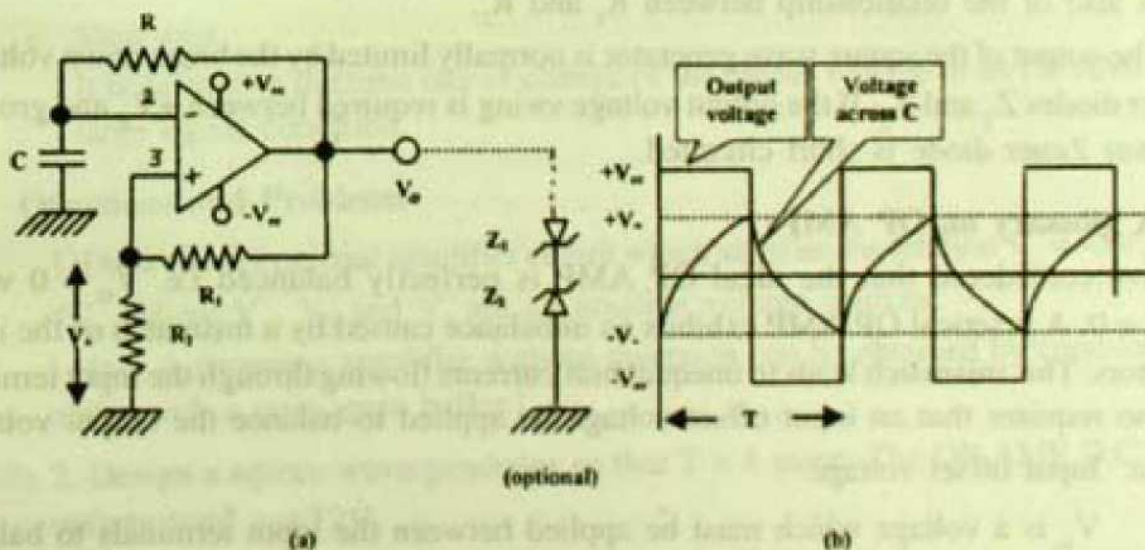


Fig. 6.16 Square wave generator : (a) circuit, (b) waveform



shows a simple OP AMP square wave generator. When the dc supply voltage  $+V_{cc}$  and  $-V_{cc}$  are switched on the input 2 is at the ground potential since the capacitor C is uncharged. However, because of the dc offset voltage at the output, a small positive or negative voltage  $V_+$  appears between the noninverting input 3 and ground. Although very small, this voltage will drive the OP AMP into saturation ( $\pm V_{cc}$ ). Assuming the offset voltage  $V_{cc}$  to be positive the voltage  $-V_+$  is also positive. Since initially the capacitor acts as a short circuit, the gain of the OP AMP is very high; hence  $V_+$  drives the OP AMP to its positive saturation the output voltage is at  $+V_{cc}$ , the capacitor C starts charging towards  $+V_{cc}$  through the resistor R. As soon as the voltage across C become slightly more positive than  $V_+$  the output of the OP AMP is forced to switch to a negative saturation  $-V_{cc}$ . At this point the voltage across  $R_2$  is also negative, since

$$V_+ = \frac{R_1}{R_1 + R_2} \cdot (-V_{cc})$$

Hence the output of the OP AMP will remain in negative saturation. This situation will continue until the capacitor discharges, and then recharges to a negative voltage slightly higher than  $-V_+$ . Now, when the capacitor voltage makes this inverting input more negative than the noninverting input, the OP AMP output, is again brought back to positive saturation at  $+V_{cc}$ . This completes one cycle. The output wave form  $V_o$  and the voltage across C is shown in Fig. 6.16 (b)

The time period of the output wave form is

$$T = \frac{1}{f} = 2RC \ln \left( \frac{2R_1 + R_2}{R_2} \right)$$

This equation shows that the output frequency  $f$  is not only a function of the time constant RC but also of the relationship between  $R_1$  and  $R_2$ .

The output of the square wave generator is normally limited by the break down voltages of zener diodes  $Z_1$  and  $Z_2$ . If the output voltage swing is required between  $+V_Z$  and ground, the lower Zener diode is short circuited.

## 6.5 A glossary on OP AMPs

We have considered that the ideal OP AMP is perfectly balanced i.e.  $V_o = 0$  when  $V_1 - V_2 = 0$ . A practical OP AMP exhibits an unbalance caused by a mismatch of the input transistors. This mismatch leads to unequal bias currents flowing through the input terminals and also requires that an input offset voltage be applied to balance the output voltage.

### a. Input offset voltage

$V_{io}$  is a voltage which must be applied between the input terminals to balance the amplifier (Fig. 6.17)



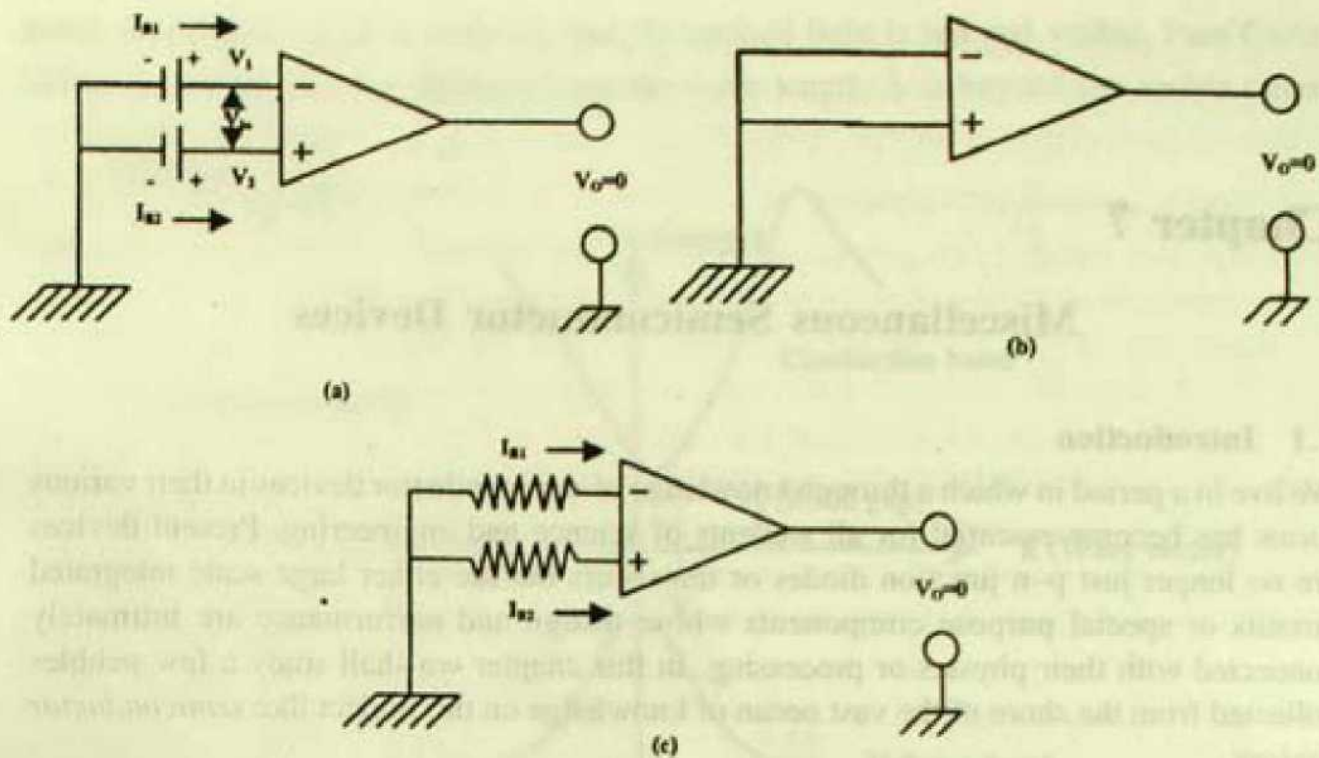


Fig. 6.17

**b. Output offset voltage**

If the two inputs are grounded we often get an output voltage known as output offset voltage (Fig. 6.17 b).

**c. Input bias current**

This is defined as one-half the sum of the separate currents entering the two input terminals of a balanced amplifier (Fig. 6.17 c).

**d. Input offset current**

It is defined as  $I_{io} = I_{B1} - I_{B2}$  (Fig. 6.17 a)

**e. Slew rate**

It is defined as the time rate of change of the output voltage in an OP AMP under large signal condition.

## 6.6 Questions and Problems

1. Draw the operational amplifier circuit which satisfies the relation  $V_o = -5V_1 + 9V_2 - 4V_3$ , where  $V_1$ ,  $V_2$  and  $V_3$  are all positive voltage sources.

[ Hint. A summing amplifier without inversion can be obtained by combining the adder with a unity gain buffer.]

2. Design a square wave generator so that  $T = 1$  msec. The OP AMP D.C supply voltage used  $= \pm 12V$ .

[ Hint. use  $R_2 = 1.16 R_1$ ].

## Chapter 7

### Miscellaneous Semiconductor Devices

#### 7.1 Introduction

We live in a period in which a through knowledge of semiconductor devices in their various forms has become essential for all students of science and engineering. Present devices are no longer just p-n junction diodes or transistors but are either large scale integrated circuits or special purpose components whose design and performance are intimately connected with their physics or processing. In this chapter we shall study a few pebbles collected from the shore of the vast ocean of knowledge on the subject like *semiconductor devices*.

#### 7.2 Light Emitting Diode (LED)

The development of optoelectronics has been controlled largely by the types of light sources available. Before the discovery of LEDs the only conveniently available light sources were tungsten and gas discharge lamps. Each source has its particular advantages. The energy released when an electron falls from the conduction to valence band appears in the form of radiation. Such devices (p-n junction) are called light emitting diode (LED). In contrast to tungsten lamps, LEDs have narrow band spectral emission, very fast response, are smaller in size, are robust mechanically, and have very long lifetime. They are generally low power devices though same types, emitting coherent radiation, are also available.

LEDs utilize the recombination of the excess carriers, produced by injection in a forward-biased p-n junction diode, to obtain light emission for display purpose. Since photons carry practically no momentum, the momentum conservation law makes direct radiative recombination impossible without the participation of several phonons that can carry the excess momentum. This makes radiative recombination a very unlikely process in indirect band materials, such as Si and Ge, the two most popular semiconductors. The recombination in such materials is mainly nonradiative, via recombination centers, and just heats the lattice thereby generating phonons. In III-V semiconductors like GaAs, InSb the band gap is direct and E-k (energy-wave vector) diagram is such that (Fig 7.1) both conduction electrons and holes have very low momentum values. Direct radiative recombination is therefore very likely and it is such semiconductors that are used for light generation. Also there are compound semiconductors like  $\text{GaAs}_x\text{P}_{1-x}$  (gallium arsenide phosphide) which has a variable band gap. The band gap depends on x, the arsenic percentage. Band gap varies from 1.43 eV for pure GaAs (x=1) to 2.26 eV for pure GaP (x=0). At x=0.44 the band gap is still



direct, as in GaAs (Gap is indirect), but the emitted light is red and visible. Pure GaAs LEDs cannot be used for displays since the wave length  $\lambda$  is beyond the visible range

$$\lambda(\mu\text{m}) \equiv \frac{1.234}{\epsilon_g(\text{eV})}$$

$\epsilon_g = 1.43 \text{ eV at } 300^\circ\text{C}$

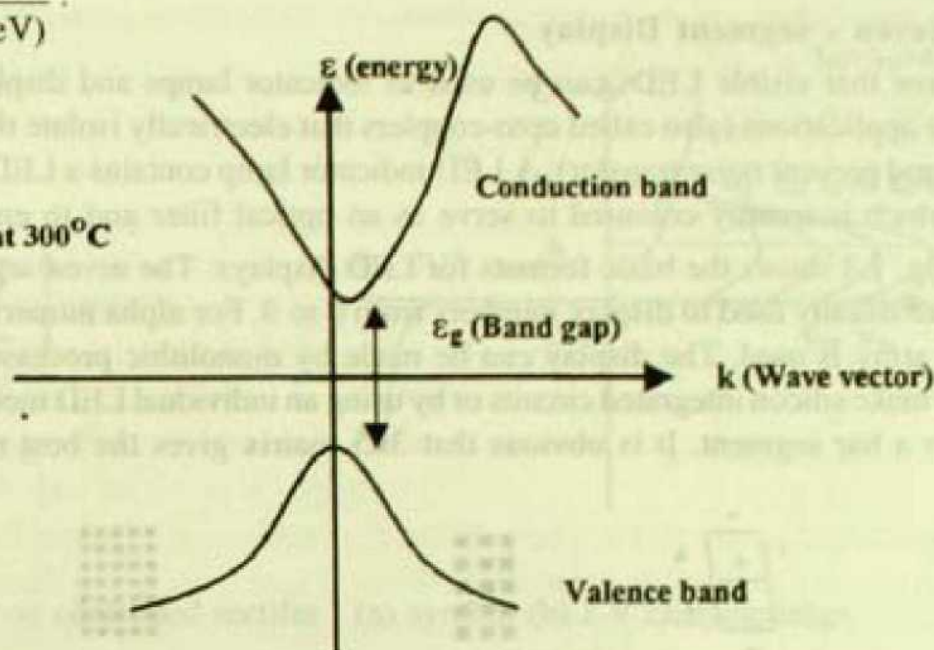


Fig. 7.1 Energy bands in momentum space for GaAs

Gap itself may also be used for LEDs, although it is an indirect band-gap material, by inclusion of special impurities in it. This gives rise to impurity levels which act as traps and make radiative recombination possible. Addition of ZnO or CdO gives red light with relatively high efficiency. Addition of nitrogen or sulphur shifts the radiation to green with much lower efficiency but the greater sensitivity of the human eye (Fig. 7.2) to green than compensates for that.

LEDs are used widely in electronic circuits. Parallel connection of many LEDs

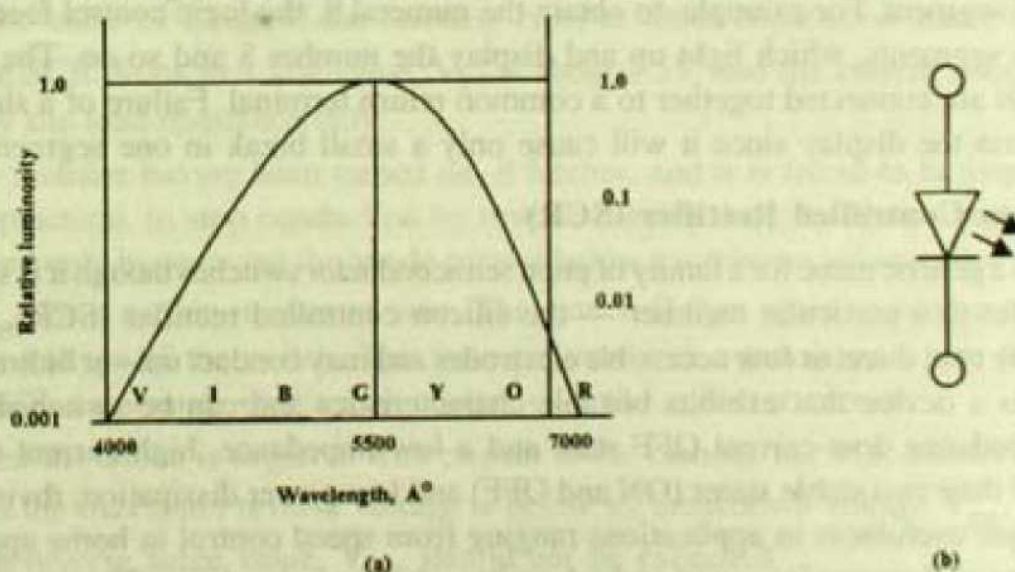


Fig. 7.2(a) Standard spectral sensitivity of human eye (b) symbol of LED



of small area may be used to display numbers and letters, as is done in modern pocket calculators, digital watches, etc. Infrared LEDs are useful in opto-isolators, and is a potential source for fibre-optics communication.

### 7.3 Seven - segment Display

We know that visible LEDs can be used as indicator lamps and displays, and for opto-isolator applications (also called opto-couplers that electrically isolate the input and output signal and prevent noise transfer). A LED indicator lamp contains a LED chip and a plastic lens, which is usually coloured to serve as an optical filter and to enhance contrast.

Fig. 7.3 shows the basic formats for LED displays. The seven segment and the  $3 \times 5$  array are usually used to display numbers from 0 to 9. For alpha numeric displays the  $5 \times 7$  matrix array is used. The display can be made by monolithic processes similar to those used to make silicon integrated circuits or by using an individual LED mounted on a reflector to form a bar segment. It is obvious that  $3 \times 5$  matrix gives the best result.

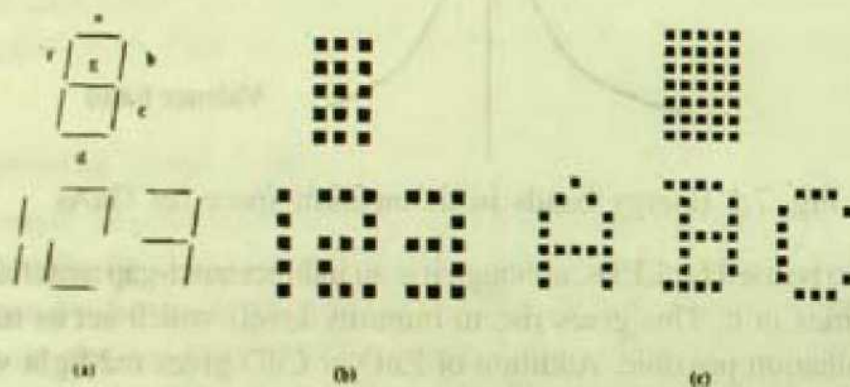


Fig 7.3 (a) The seven segment for obtaining numerical display; (b)  $3 \times 5$  array (numeric); (c)  $5 \times 7$  LED matrix for alphanumeric display.

A digital control system (discussed in next chapter) is used to direct the current to the desired segment. For example, to obtain the numeral 8, the logic control feeds current into all the segments, which light up and display the number 8 and so on. The cathodes of the LEDs are connected together to a common return terminal. Failure of a single LED does not ruin the display since it will cause only a small break in one segment.

### 7.4 Silicon Controlled Rectifier (SCR)

Thyristor is a generic name for a family of pnpn semiconductor switches though it is sometimes used to refer to a particular member — the silicon controlled rectifier (SCR). Different devices have two, three, or four accessible electrodes and may conduct uni- or bidirectionally. Thyristor is a device that exhibits bistable characteristics and can be switched between a high impedance, low-current OFF state and a low-impedance, high-current ON state. Because of their two stable states (ON and OFF) and low power dissipation, thyistors have found unique usefulness in applications ranging from speed control in home applications to switching and power inversion in high-voltage transmission lines. Thyristors are now available with current ratings from a few mA to over 5000 A and voltages over 10,000



V. The most commonly encountered thyristor is the silicon controlled rectifier (SCR). The current voltage characteristic and circuit symbol of an SCR are shown in Fig. 7.4. This is a special type of diode which, in order to start conduction, must have not only a positive

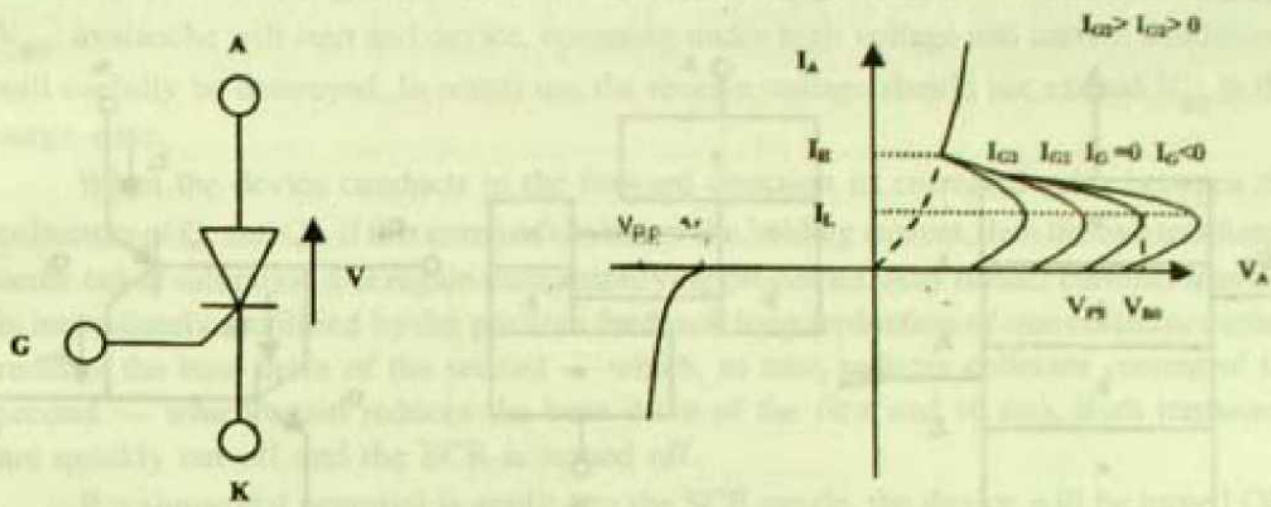


Fig. 7.4 Silicon controlled rectifier : (a) symbol (b) I-V characteristics

anode-cathode voltage  $V$  but also a high enough positive pulse applied to a third electrode called the gate  $G$ .

If a positive voltage, smaller than some maximum  $V_{FS}$  (F refers to forward, S to surge), is applied between the anode  $A$  and cathode  $K$ , the SCR passes only a leakage current, smaller than some specified value. However, if a pulse of certain minimum voltage and current is applied as a trigger  $I_G$  between gate and cathode, the SCR switches to conduction even if  $V_{AK}$  is just a few volts positive. The device also switches to conduction if its forward voltage exceeds a value called the *break-over voltage*  $V_{BO}$ . At  $V_{BO}$  the leakage current reaches the value of  $I_L$  called the *latching current*. Once conduction starts, the voltage across the SCR drops to a low value,  $V_T$ , at about 1.5V, and the current rises to a value limited by the load resistance only.

The rectifier having been turned on, it latches, and it is found to be impossible, or rather impractical, to stop conduction by reverse biasing the gate. However, turn-off can be achieved only by reducing the anode current below a minimum value, called the *holding current*  $I_H$ , for a certain minimum time. The gate will then again assume control of the breakover voltage of the switch. This is in contrast with a transistor, where the base (or gate in FET) commands continuous control over the collector current.

When the anode is negative with respect to the cathode the SCR blocks conduction as long as the maximum reverse voltage is below its breakdown voltage  $V_{BR}$ . In use, the maximum reverse surge value,  $V_{RS}$ , should not be exceeded.

The physical explanation for SCR operation is as follows. The device is a four layer



structure with three junctions  $J_1, J_2, J_3$ , as shown in Fig. 7.5. The structure may be viewed as a combination of two transistors npn and pnp connected back to back. The significance of this view is that it shows the current gain of the sections, and existence of positive feedback. These two factors are necessary to obtain the desired switching and latching operation.

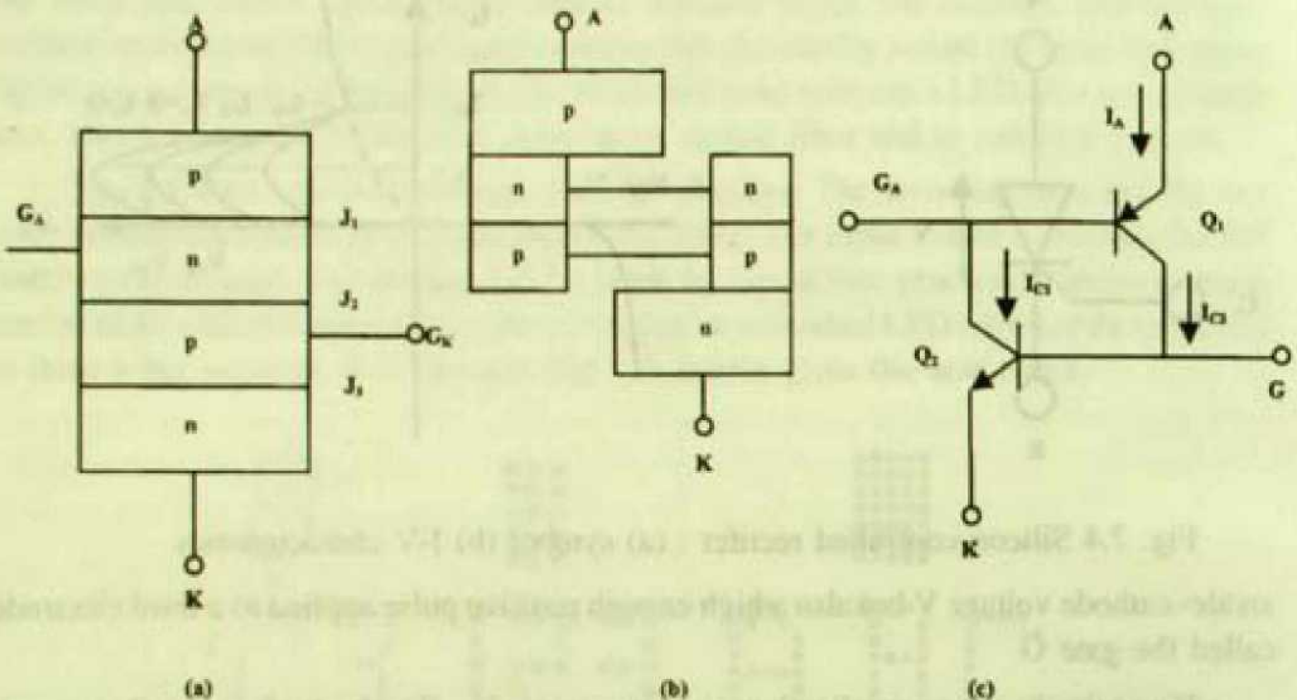


Fig. 7.5 Two-transistor model of SCR : (a) cross-section, (b) division into two transistors, (c) two-transistor model

Application of a positive voltage between anode and cathode will not help in conduction because the junction  $J_2$  will be under reverse bias and hence blocking. But looking at Fig. 7.5 (c) we find that  $T_2$  is the collector base junction in both equivalent transistors  $Q_1$  and  $Q_2$  and reverse voltage across it is normal to transistor operation in the active region. Current can therefore flow in  $Q_1$  if a base current is supplied i.e. if a current pulse is fed into the gate  $G_K$  which serves as the base of  $Q_1$ . The resulting collector current of  $Q_1$  serves as the base drive for  $Q_2$ , which also starts conduction then. The collector current of  $Q_2$  serves as additional base drive for  $Q_1$  and takes over when the original gate trigger pulse ends. This constitutes a positive feedback loop, and the two transistors very quickly become heavily saturated and the device turns on. In saturation the junction  $J_2$  become forward biased (like any other collector junction of a saturated BJT) so that all three junctions  $J_1, J_2$  and  $J_3$  are now forward biased. The total forward voltage drop  $V_T$  of the SCR consists of the algebraic sum of these voltages in addition to the ohmic voltage drop in the bulk of the SCR. However,  $J_1$  and  $J_3$  contribute a voltage of opposite sign to that contributed by  $J_2$ , resulting in a total of  $V_T$  between 1 to 2 volts depending on the current.



At forward voltages higher than  $V_{BO}$ , the increased leakage current switches the device ON even without gate triggering. When the anode-cathode voltage is negative, both  $J_1$  and  $J_3$  are reverse biased, and since they are also the emitter-base junctions of the two-transistor model the device remains cut off. If the reverse voltage exceeds the breakdown voltage  $V_{BR}$ , avalanche will start and device, operating under high voltage and current conditions, will usefully be destroyed. In actual use the reverse voltage should not exceed  $V_{RS}$  in the surge case.

When the device conducts in the forward direction its current divides between the collectors of  $Q_1$  and  $Q_2$ . If this current falls below the holding current, then the two transistors come out of saturation and regain their amplifying properties. Any further current reduction is immediately amplified by the positive feedback loop (reduction of one collector current reduces the base drive of the second — which, in turn, reduces collector current of the second — which again reduces the base drive of the first and so on). Both transistors are quickly cut off and the SCR is turned off.

If a sinusoidal potential is applied to the SCR anode, the device will be turned OFF once each alternate half cycle (i.e. when the voltage falls below the holding voltage) provided it is triggered ON regularly. The average rectified current can be varied over wide limits by controlling the point in each half cycle at which the SCR is turned ON. This is called phase control and is widely used in controlling ac power, for example, for heating, lighting, motor drive, electric welding, and a variety of other industrial control applications.

### 7.5 Diac and Triac

The diac (diode ac switch) and triac (triode ac switch) are bidirectional thyristors. They have ON and OFF states for positive or negative anode voltages and are therefore useful in ac applications.

The diac has mainly two structures — (i) the ac trigger diode and (ii) the bilateral p-n-p-n diode switch. The former is simply a three layer device similar in construction to a bipolar transistor, except that doping concentrations at the two junctions are approximately the same to have a symmetrical, bidirectional characteristics (Fig. 7.6). No contact is made to the base region. When a voltage of any polarity is applied to a diac,

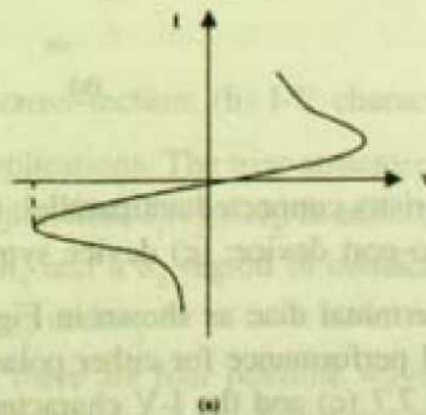


Fig. 7.6 (a) Typical characteristics of a diac



one junction will be forward biased and the other reverse biased. The current will obviously be limited by the leakage current of the reverse biased junction. At sufficiently large reverse bias, breakdown occurs at  $V_{BR}(1-\alpha)^m$ , where  $V_{BR}$  is the avalanche breakdown voltage of the p-n junction,  $\alpha$  is the common base current gain, and  $m$  is a constant. As the current increases after breakdown,  $\alpha$  increases, causing a reduction in the terminal voltage. This reduction gives rise to a negative differential resistance region.

However, in practice a diac is fabricated from p-n-p-n thyristors for its greater efficiency and higher break over voltages. The bidirectional p-n-p-n diode switch behaves like two conventional Shockley diodes connected antiparallel to permit the accommodation to voltage signals of two polarities, as in Fig. 7.7 (a), where MT1 stands for main terminal 1, and MT2 for the main terminal 2. Using the short-cathode principle, we can integrate this

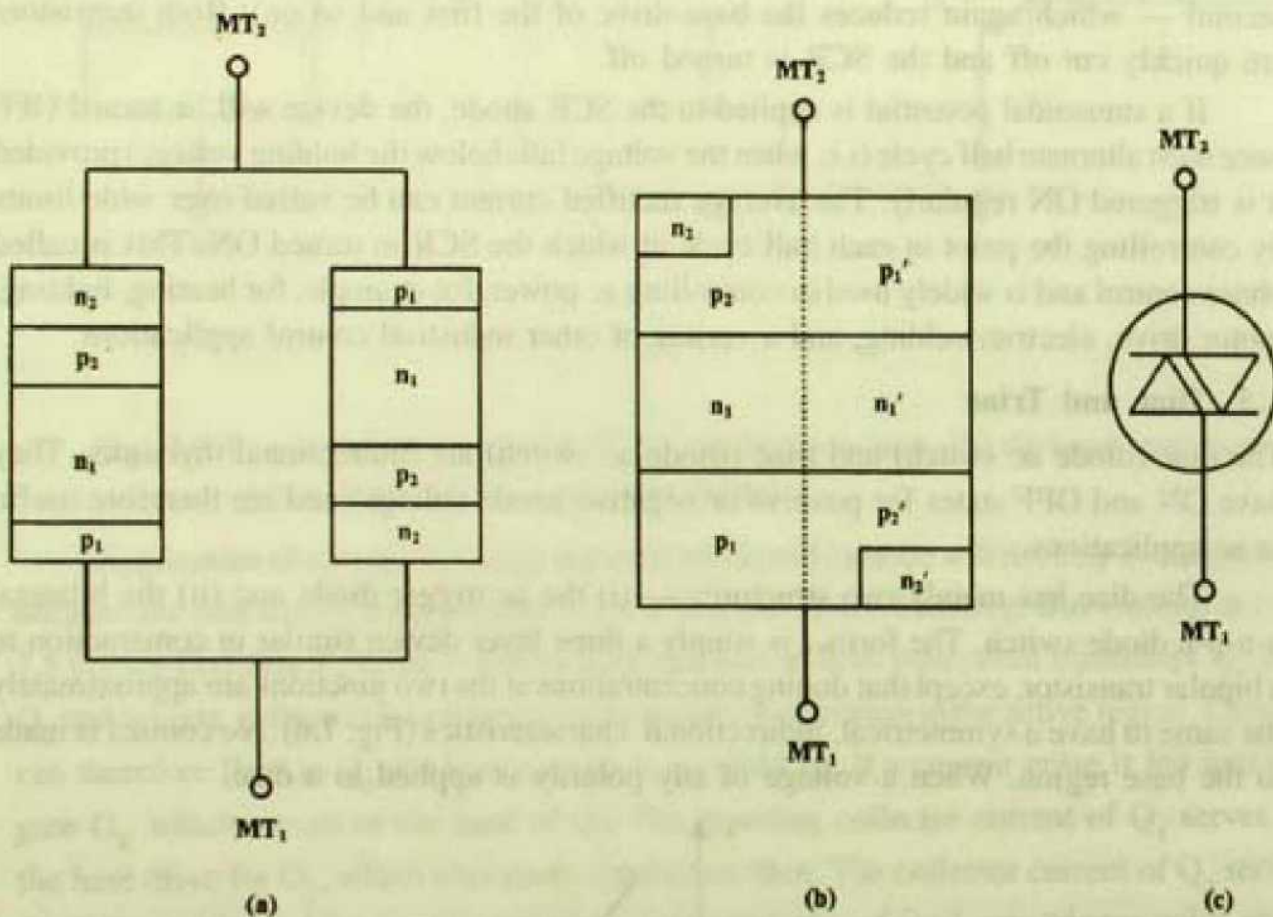


Fig. 7.7 Diac : (a) two thyristors connected antiparallel; (b) integration of the diodes into a single two-port device; (c) device symbol

arrangement into a single two-terminal diac as shown in Fig. 7.7 (b). The symmetry of this structure results in identical performance for either polarity of applied voltage. The circuit symbol is shown in Fig. 7.7 (c) and the I-V characteristics in Fig. 7.6. The diac can be triggered into conduction by exceeding the breakover voltage. Because of its



regenerative action by positive feedback, the bidirection p-n-p-n diode switch has a larger negative resistance and smaller forward voltage drop than that of an ac trigger diode.

The triac is used for current switching in ac circuits. By two SCRs, ac power can also be controlled, since it is a very common requirement, a device consisting of two SCRs connected in inverse parallel has been developed. This is a triac. The triac can switch the current in either polarity between the gate and one of the two main terminals (Fig. 7.8 a). The triac is very useful in motor speed control, temperature control,

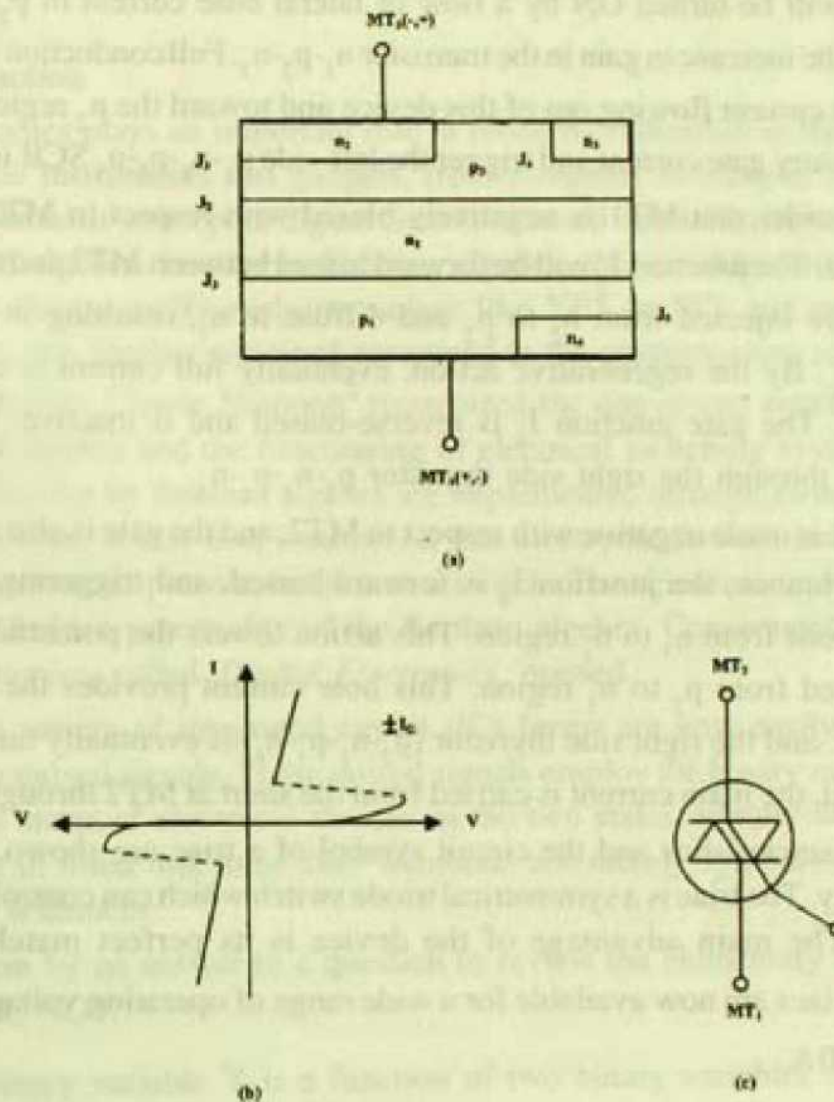


Fig. 7.8 Triac : (a) cross-section; (b) I-V characteristics; (c) symbol

light dimmer, and other applications. The triac structure is considerably more complicated than an SCR due to its five junctions  $J_1, J_2, \dots, J_5$ . In addition to the basic  $p_1-n_1-p_2-n_2$  structure, there are a junction gate  $n_3$  and a  $n_4$  region in contact with  $MT_1$ . It may be noted that  $p_1$  is shorted to  $n_4$ ,  $p_2$  to  $n_2$  and  $n_3$ .

It is easily seen that there are four possible ways to trigger a triac : two polarities of overall voltage (i.e.  $MT_1$  is either + or -) and two possible gate currents (+ -). When the main terminal  $MT_1$  is + with respects to  $MT_2$  and a positive voltage (or current) is



applied to the gate (also with respect to MT2), the device behavior is identical to that of an SCR. The junction  $T_4$  is reverse biased and is nonconducting : the gate current is supplied through the gate close to  $n_3$  region; since junction  $T_5$  is also reverse biased and hence inactive, the main current flows through the left side of the  $p_1-n_1-p_2-n_2$  section.

Now let a negative voltage be applied to the gate in the above case. The junction  $J_4$  will now be forward biased and electrons will be injected from  $n_3$  to  $p_2$ . The auxiliary SCR  $p_1-n_1-p_2-n_2$  will be turned ON by a flow of lateral base current in  $p_2$  toward the  $n_3$  region because of the increase in gain in the transistor  $n_1-p_2-n_3$ . Full conduction of this auxiliary SCR results in the current flowing out of this device and toward the  $n_2$  region. This current provides the necessary gate current and trigger the left-side  $p_1-n_1-p_2-n_2$  SCR into conduction.

We now consider that MT1 is negatively biased with respect to MT2, and the gate is positively biased. The junction  $J_3$  will be forward biased between MT2 and the gate shorted to  $p_2$ . Electrons are injected from  $n_2$  to  $p_2$  and diffuse to  $n_1$ , resulting in an increase of forward bias of  $J_2$ . By the regenerative action, eventually full current is carried through the short at MT2. The gate junction  $J_4$  is reverse-biased and is inactive. The full direct current is carried through the right side thyristor  $p_2-n_1-p_1-n_4$ .

Finally, MT1 is made negative with respect to MT2, and the gate is also made negative. Under this circumstance, the junction  $J_4$  is forward biased, and triggering is initiated by injection of electrons from  $n_3$  to  $n_1$  region. This action lowers the potential at  $n_1$ , causing holes to be injected from  $p_2$  to  $n_1$  region. This hole current provides the base drive for  $p_2-n_1-p_1$  transistor, and the right side thyristor ( $p_2-n_1-p_1-n_4$ ) is eventually turned ON. Since  $J_3$  is reverse biased, the main current is carried from the short at MT2 through the  $n_4$  region.

The  $I-V$  characteristics and the circuit symbol of a triac are shown in Fig. 7.8 (b) and (c) respectively. The triac is a symmetrical triode switch which can control loads supplied with ac power. The main advantage of the device is its perfect matching of output characteristics. Triacs are now available for a wide range of operating voltages ( $= 2000V$ ) and currents  $= 300A$ .



## Chapter 8

### Digital Electronics

#### 8.1 Introduction

Digital electronics plays an important role in modern civilization in the sense that the vast majority of the instruments and gadgets, from computer to camera, are based on digital electronics. The basic concept of digital circuits is based on the fundamental work of George Boole<sup>1</sup>. In the algebra, known as Boolean algebra, the variables are allowed to assume only *two possible mutually exclusive values* like YES or NO, and never like MAY BE. Unfortunately, this algebra remained unnoticed to the contemporary scientific community for about a century. Claude Shannon<sup>2</sup> recognized the one-to-one correspondence between these form of algebra and the functioning of electrical switching systems. The reasoning processes called for by Boolean algebra are implemented through switches, acting as *logic gates*. By the time, it was also established that electronic devices (such as valves at that time and transistors at present, behave as a switch. The scientists and engineers therefore exploited the hidden potentiality of the Boolean algebra. Consequently, a new era in the field of electronics, called *Digital Electronics*, opened.

A great variety of integrated circuit (IC) forms are now easily available for logic operations on pulsed signals. These pulsed signals employ the binary number system, using OFF and ON states of electronic devices as the two states, usually written as 1 and 0. It must be kept in mind that these two 'numbers' are merely symbols and not numbers in the sense of arithmetic.

We begin by an answer to a question to review the elementary idea of basics gates — AND, OR, NOT.

Q. A binary variable Y is a function of two binary variables A and B, i.e.

$$Y = f(A, B)$$

In how many ways Y can be defined ? Put them in truth table from. Interpret the meaning of sponce of the forms of Y.

Ans. 16 ways.

1. George Boole (1815-18), *The Mathematical Analysis of Logic*, Cambridge University Press, London, 1847. *ibid*, *On Investigation of the Laws of Thought*, London, 1857.

2. Claude Shannon, *A Symbolic Analysis of Relay and Switching Circuits*, AIEE vol. 157, p 713-723, Dec. 1938.



Since A and B may each assume only two values, there are only four possible combinations of the two variables. It is therefore feasible to define the function  $Y = f(A, B)$  by specifically stating the values of Y for each of the 4 combination of A and B. There are  $2^4(=16)$  possible functions of 2 variables and are specified in tabular form in Table 8.1.

**Table 8.1** Sixteen possible Boolean functions of two variables

A	B	$f_0$	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	$f_6$	$f_7$	$f_8$	$f_9$	$f_{10}$	$f_{11}$	$f_{12}$	$f_{13}$	$f_{14}$	$f_{15}$
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

It can be readily remarked that  $f_0$  and  $f_{15}$  are not functions at all, being independent of A and B.

$f_3 = A$  and do't care for B

$f_5 = B$  and do't care for A

$f_{10} = \bar{B} = \bar{f}_5$

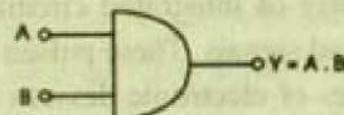
$f_{12} = \bar{A} = \bar{f}_3$

$f_1 \Rightarrow$  This function can be characterized in words by saying that "Y is 1 if and only if A is 1 *and* also B is 1". Hence this function is referred to as the **AND** function and the fundamental relationship is written as

$Y = A \text{ and } B,$

denoted by  $Y = A.B$

with the symbol shown.



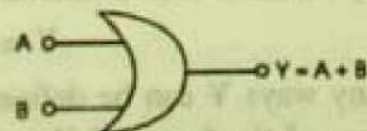
It is also known as multiplication logic, since it obeys the arithmetic laws of multiplication between the values of A and B.

$f_7 \Rightarrow$  This table defines the **OR** operation. This can be stated in the form : "Y is 1 if A is 1 **OR** B is 1". This is sometimes called *inclusive OR* because it includes "Y is 1 if both A and B are 1". It is written as,

$Y = A \text{ or } B,$

denoted by  $Y = A+B$

with the symbol shown.

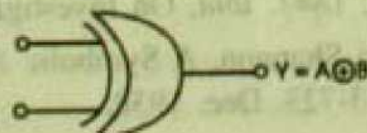


$f_6 \Rightarrow$  This table defines **exclusive OR** (also written as X- OR) operation. In words, "Y is 1 if A is 1 **OR** if B is 1 provided that this condition of truth is *exclusive*, that is, A and B are not 1 simultaneously. Mathematical notation for X- OR is

$$Y = A \oplus B$$

$$= A\bar{B} + \bar{A}B$$

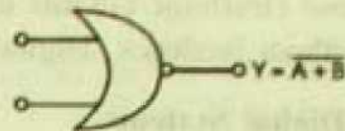
with the symbol shown.





$f_8 \Rightarrow \bar{f}_7$ . It is seen that this table defines the complement of OR and is said to be NOR. This can be expressed as

$Y = \overline{A + B}$  with the symbol shown.



$f_9 = \bar{f}_6$  - X-NOR and  $Y = \overline{A \oplus B}$ , known as "equality detector".

$f_{14} = \bar{f}_1$  - NAND and  $Y = \overline{A \cdot B}$  with the symbol  $Y = \overline{A \cdot B}$

$f_4 = \bar{f}_{11}$ , where  $f_{11} = A \supset B$  means "A implies B"  $= \bar{A} + B$

$f_2 = \bar{f}_{13}$ , —  $f_{13} = B \supset A$  means "B implies A"  $= A + \bar{B}$

It may be remarked that AND and OR gates or its complements or combinations may assume any number of inputs.

**Example 8.1** Prove that the following all NAND circuit is equivalent to ex-OR.

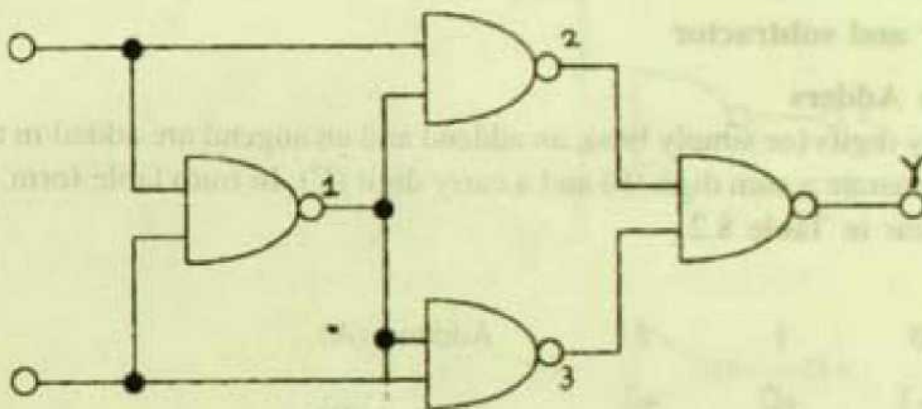


Fig. for example 8.1

**Solution :** Point 1 is  $\overline{A \cdot B}$

Point 2 is  $\overline{A \cdot \overline{A \cdot B}} = \overline{A + AB} = \overline{A + B}$

Point 3 is  $\overline{B \cdot \overline{A \cdot B}} = \overline{B + AB} = \overline{B + A}$

Hence  $Y = \overline{(\overline{A \cdot \overline{A \cdot B}}) \cdot (\overline{B \cdot \overline{A \cdot B}})}$

$$= \overline{(\overline{A + B}) \cdot (\overline{A + B})}$$

$$= \overline{\overline{A + B} + \overline{A + B}}$$

$$= A \cdot \bar{B} + \bar{A} \cdot B$$

$$= A \oplus B$$

A	B	$A \oplus B$
0	0	0
0	1	1
1	0	1
1	1	0

## Digital Systems

Digital systems are those electronic circuits in which we use the three basic logic gates many times with or without feedback. Digital systems are classified into two classes:

### A. Combinational Digital Systems

In this system the outputs at a given instant of time depends only upon the values of the inputs at the same moment. Even a ROM (Read Only Memory) is a combinational system. The memory of a ROM refers to the fact that it *memorizes* the fundamental relationship between the output variables and the input variables. It does not store bits of information.

### B. Sequential Digital Systems

In such systems the output is obtained in time sequence of a clock pulse. Under these circumstances a sequential circuit must possess a memory i.e. it will *remember* its earlier logic state (either 0 or 1) immediately before the arrival of a clock pulse. These systems are constructed with heavy positive feedback.

## 8.2 Combinational Circuits

### (a) Adder and subtractor

#### A. Binary Adders

Two binary digits (or simply bits), an addend and an augend are added in the manner of Fig. 8.1, to generate a sum digit (S) and a carry digit (C). In truth table form, the results of Fig. 8.1 appear in Table 8.2.

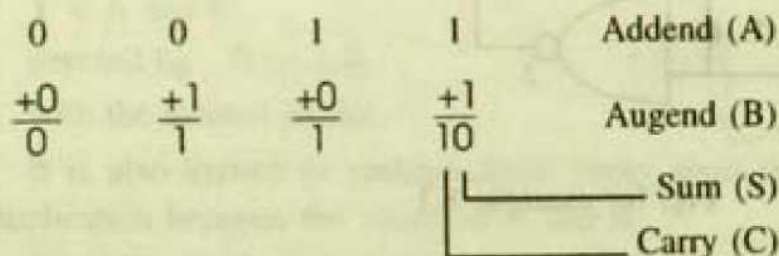


Fig. 8.1 Binary addition principle

Table 8.2

A	B	S	C
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

The carry digit is seen to be related to the addend and augend by AND logic

$$C = A.B$$

(8.1)



while the sum is given by the X-OR operation

$$S = A \oplus B = \overline{A}B + A\overline{B} \quad (8.2)$$

A gate structure (Fig. 8.2) which performs such an addition of 2 bits is called a half adder. Fig. 8.2(a) corresponds directly to the equation (8.1) and (8.2), (b) is the symbol and (c) is a NAND only circuit.

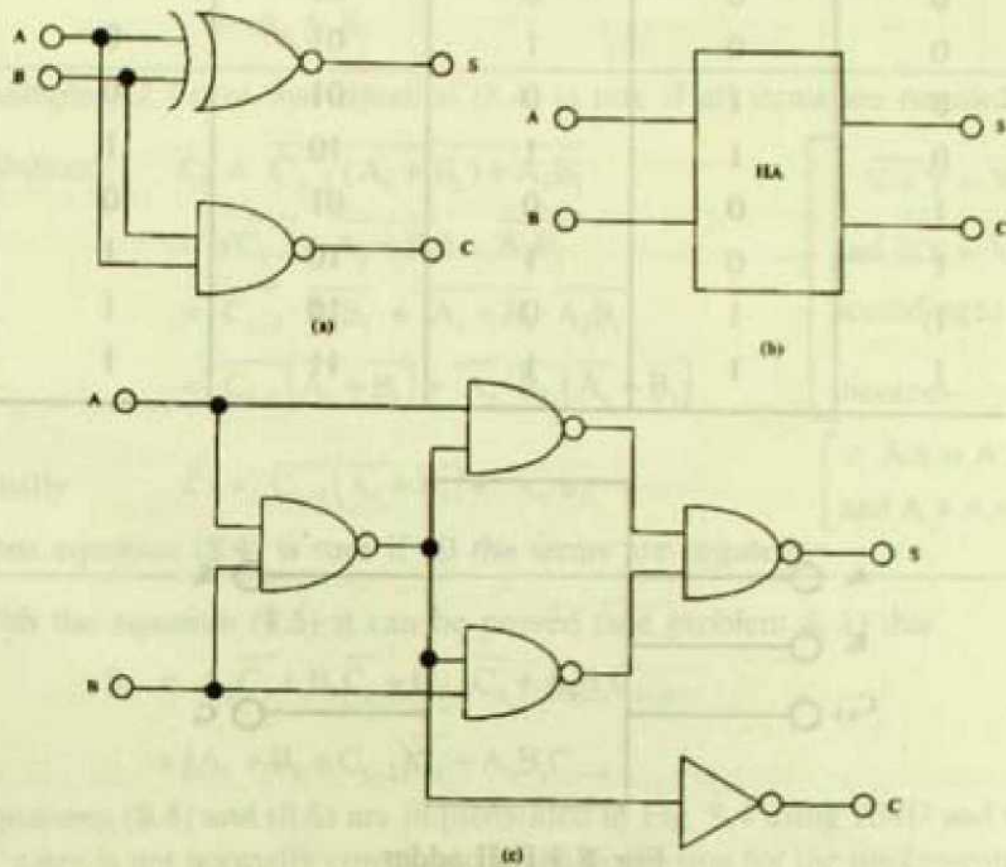


Fig. 8.2 Half adder

### Full Adder

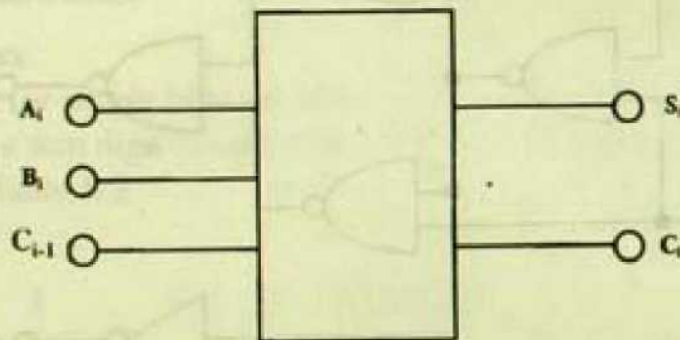
The carry bit needs to be added as soon as the number of bits in the binary number exceeds one. Two binary N-bit numbers can be added as shown below:

		$C_{N-2}$	$C_1$	$C_0$			
A	=	$A_{N-1}$	$A_2$	$A_1$	$A_0$		
B	=	$B_{N-1}$	$B_2$	$B_1$	$B_0$		
<hr/>							
S	=	$C_{N-1}$	$S_{N-1}$	$C_2$	$S_2$	$S_1$	$S_0$

A full adder is essentially a half adder with a carry input signal. Fig. 8.3 shows the symbol and truth table (8.3) for this operation.

**Table 8.3**

Inputs			Outputs	
Carry $C_{i-1}$	Addend $A_i$	Augend $B_i$	Sum $S_i$	Carry $C_i$
0	0	0	00	0
0	0	1	01	0
0	1	0	01	0
0	1	1	10	1
1	0	0	01	0
1	0	1	10	1
1	1	0	10	1
1	1	1	11	1



**Fig. 8.3 Full adder**

From the truth table it is seen that

$$\begin{aligned}
 S_i &= \overline{C_{i-1}} \overline{A_i} B_i + \overline{C_{i-1}} A_i \overline{B_i} + C_{i-1} \overline{A_i} \overline{B_i} + C_{i-1} A_i B_i \\
 &= \overline{C_{i-1}} (A_i \oplus B_i) + C_{i-1} (\overline{A_i \oplus B_i}) \\
 &= C_{i-1} \oplus A_i \oplus B_i \quad (8.3)
 \end{aligned}$$

$$\begin{aligned}
 \therefore \overline{A \oplus B} &= \overline{AB + \overline{A} \overline{B}} \\
 &= \overline{AB} \cdot \overline{\overline{A} \overline{B}} \\
 &= (\overline{A} + B) \cdot (A + \overline{B}) \\
 &= \overline{AB} + AB
 \end{aligned}$$

$$C_i = \overline{C_{i-1}} A_i B_i + C_{i-1} \overline{A_i} B_i + C_{i-1} A_i \overline{B_i} + C_{i-1} A_i B_i$$



$$= C_{i-1}(A_i \oplus B_i) + A_i B_i$$

$$= C_{i-1}(A_i + B_i) + A_i B_i \quad (8.4)$$

since, in the particular case we can add  $A_i B_i$  as many times as we wish ( $A+A=A$ ).

$$\begin{aligned} A_i \oplus B_i + A_i B_i &= A_i \oplus B_i + A_i B_i + A_i B_i \\ &= A_i \bar{B}_i + \bar{A}_i B_i + A_i B_i + A_i B_i \\ &= A_i + B_i \end{aligned}$$

**Example 8.2** Prove that equation (8.4) is true if all terms are negated.

**Solution :**  $\bar{C}_i = \overline{C_{i-1}(A_i + B_i) + A_i B_i}$

$$= \overline{(C_{i-1} + A_i + B_i) \cdot A_i B_i}$$

$$= \overline{C_{i-1} \cdot A_i B_i + A_i + B_i \cdot A_i B_i}$$

$$= \overline{C_{i-1}(A_i + B_i) + A_i \cdot B_i (A_i + B_i)}$$

$$\because \overline{X+Y} = \bar{X} \cdot \bar{Y}$$

$$\text{and } \overline{XY} = \bar{X} + \bar{Y}$$

according to De Morgan's theorem

Finally  $\bar{C}_i = \overline{C_{i-1}(A_i + B_i) + A_i \cdot B_i}$

$$\left[ \begin{array}{l} \because AA = A \\ \text{and } A + A = A \end{array} \right] \quad (8.5)$$

Thus equation (8.4) is true if all the terms are negated.

With the equation (8.5) it can be proved (see problem 4-1) that

$$S_i = A_i \bar{C}_i + B_i \bar{C}_i + C_{i-1} \bar{C}_i + A_i B_i C_{i-1} \quad (8.6)$$

$$= (A_i + B_i + C_{i-1}) \bar{C}_i + A_i B_i C_{i-1}$$

Equations (8.4) and (8.6) are implemented in Fig. 8.4 using AND and OR gate (use of NOT gates is not normally considered in the discussion for the implementation of logic gates).

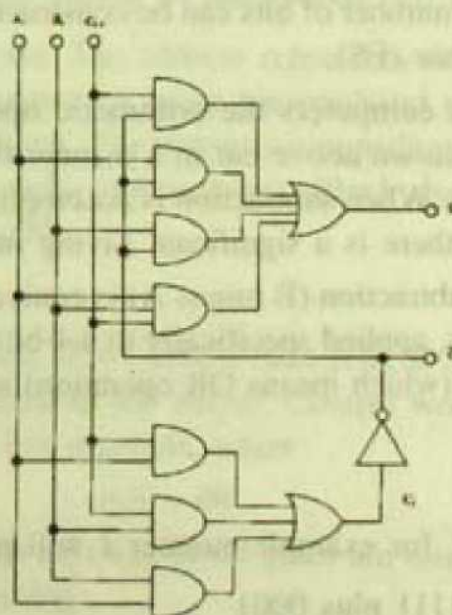


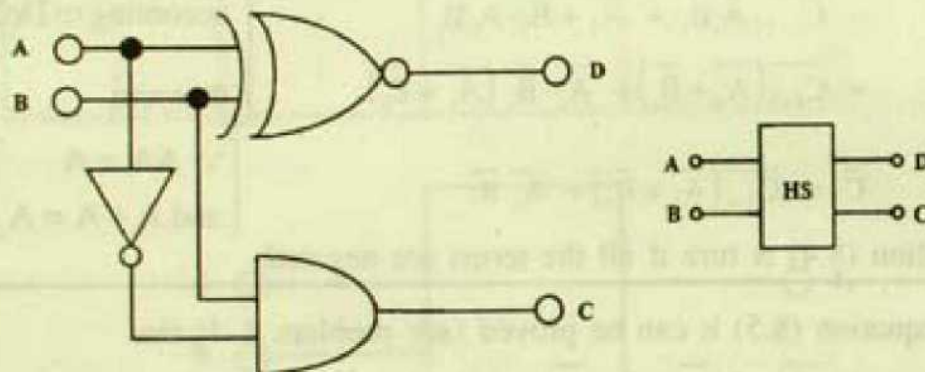
Fig. 8.4 i-th stage of a Full Adder (FA)

## Subtraction or Binary Subtractor

The rules of subtraction of binary number in any given order are specified by the truth table 8.4. A gate structure which accepts A and B as inputs and yields D and C as outputs

**Table 8.4** Truth table for subtraction

A	B	Difference D	Borrowed C
0	0	0	0
0	1	1	1
1	0	1	0
1	1	0	0



**Fig. 8.5** logic structure of subtraction

is called a *half-subtractor* (HS). It can be easily verified that D and C are given by

$$D = A \oplus B \quad \text{and} \quad C = \bar{A}B$$

Subtraction of higher number of bits can be constructed by simple logical extensions and are called full subtractor (FS).

In actual practice, in computers the arithmetic operation of subtraction is never performed in the manner shown above but in a manner that involves Adder and the use of complementary numbers. When subtraction is accomplished in this process and explicit subtractor is not required there is a significant saving in hardware.

That the process of subtraction (B minus A) is equivalent to addition can be justified by the following arguments, applied specifically to a 4-bit number. The word *plus* (*minus*) will be used in place of + (which means OR operation) and — in the present discussion to avoid any confusion.

$$A \text{ plus } \bar{A} = 1111,$$

since A is a 4-bit number; for example number 1 will mean 0001. Hence

$$\begin{aligned} A \text{ plus } \bar{A} \text{ plus } 1 &= 1111 \text{ plus } 0001 \\ &= 10000 \end{aligned}$$



Therefore  $A = 10000$  minus  $A$  minus 1.

Finally,

$$B \text{ minus } A = (B \text{ plus } \bar{A} \text{ plus } 1) \text{ minus } 10000.$$

This equation implies that to subtract a 4-bit number  $A$  from a 4-bit number  $B$  it is only required to add  $B$ ,  $\bar{A}$  and 1 ( $2^0$  bit). The term minus 10000 signifies that the addition of  $B$ ,  $\bar{A}$  and 1 may result in a fifth bit which must be ignored.

### b. Multiplexer

Multiplex means *many into one*. The function performed by a *multiplexer* (MUX) is to select 1 out of  $N$  input data sources and to transmit the selected data to a single information channel with the help of control signals. The  $n$ -position switch connected as in Fig. 8.6 (a) is the

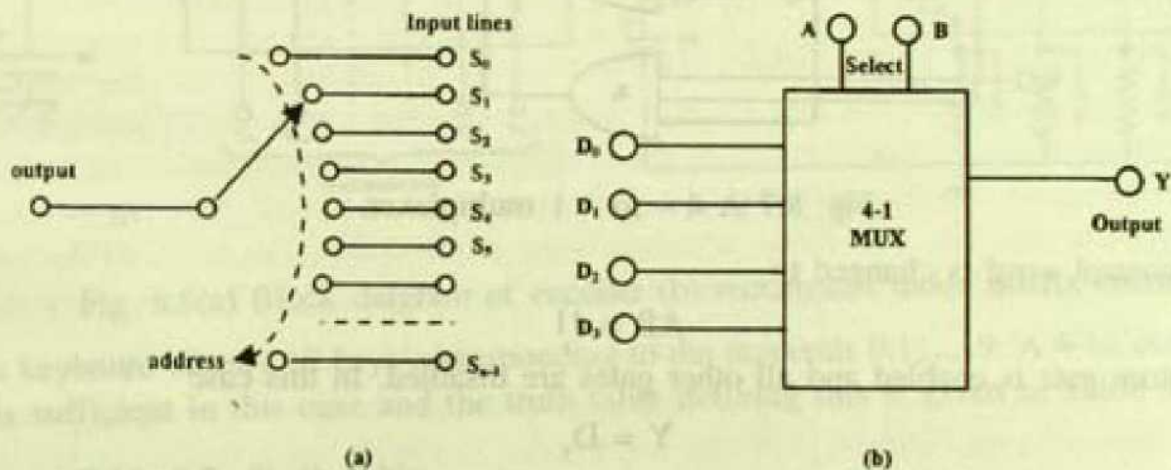


Fig. 8.6 (a) Mechanical analog of mutliplexer, (b) Block diagram

mechanical analog of a multiplexer. The address represents the angle of rotation of the switch. *Frequency multiplexer* transmits many narrow-band signals simultaneously in different portions of a frequency band. The transmission medium may be a telephone line, a coaxial cable, or space (as is radio propagation). Block diagram of a simple 4-to-1 multiplexer is shown in Fig. 8.6.(b)

### Data Selection

Fig. 8.7 shows a 4-to-1 multiplexer, also called a *data selector*. Input data bits are  $D_0$  to  $D_3$ . Only one of these is transmitted to the output. Control word  $AB$  determines which data bit is passed to the output. For example, when

$$AB = 00$$

the upper AND gate is enabled but all other AND gates are disabled. Therefore, data bit  $D_0$  is transmitted to the output giving

$$Y = D_0$$

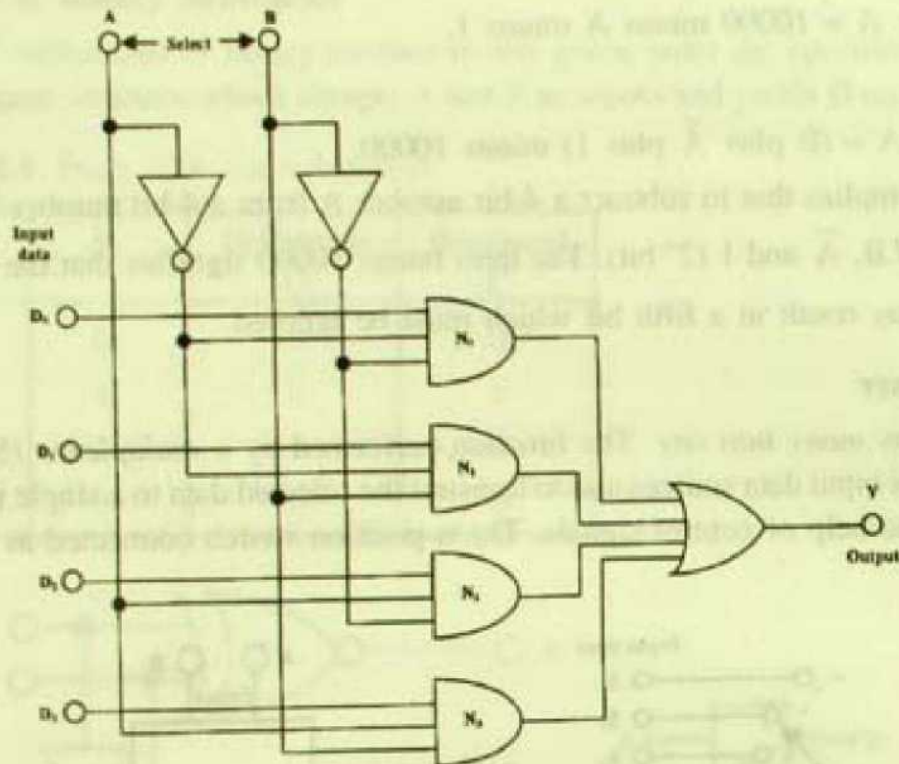


Fig. 8.7 A 4 – to – 1 multiplexer.

If the control word is changed to

$$AB = 11$$

the bottom gate is enabled and all other gates are disabled. In this case

$$Y = D_3$$

Thus we see that for  $D_2$ , for example, to be selected we must have  $AB = 10$ , which enables  $N_2$  and  $D_2$  to appear at the output  $Y$ .

### Demultiplexer

A *demultiplexer* performs the inverse process of a multiplexer. A multipole rotary switch (in Fig. 8.6 (a) the output and input words are simply interchanged) is the mechanical analog of demultiplexer.

### c. Encoder

The encoding process transforms the desired information into binary codes. Like multiplexer it has many input lines but the output lines are coded patterns which identify *each of the inputs*. In general a total of  $2^N - 1$  input lines can be represented by at best  $N$ -bit binary words.

Consider, for example, the case of a computer keyboard. It has 26 lower case and 26 capital letters, 10 numerals, and about 22 special characters (such as punctuations, symbols of arithmetic operations, and other special commands) on such a keyboard so that the total number of codes necessary is 84. Therefore a minimum of 7 bits (since  $2^6 < 84 < 2^7$ ) will



be required at the output. It should be remembered that input keys have only two possible states, pressed (logic 1) or not pressed (logic 0). The logic diagram of an encoder is given in Fig. 8.8 (a).

To illustrate the design procedure for constructing an encoder, let us consider

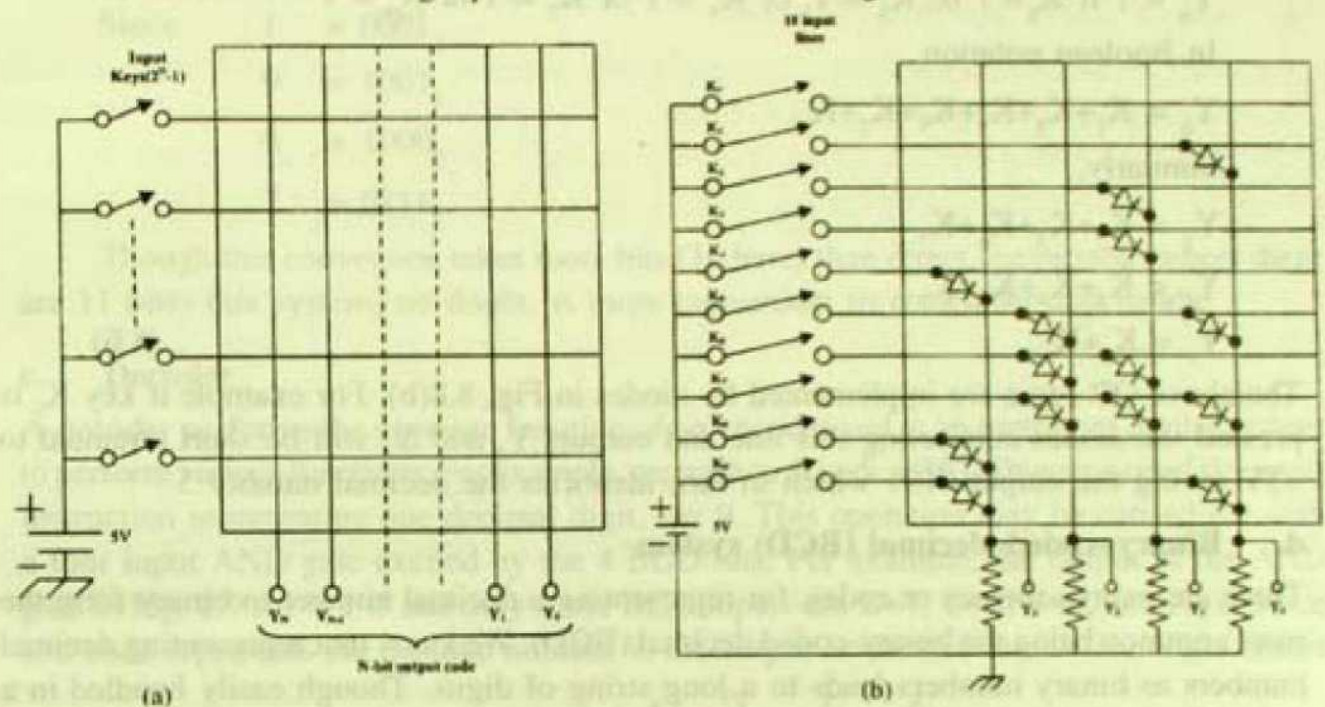


Fig. 8.8(a) Block diagram of encoder (b) rectangular diode matrix encoder.

a keyboard of only 10 keys corresponding to the numerals 0,1,.....9. A 4-bit output code is sufficient in this case and the truth table defining this is given in Table 8.5. Input

Table 8.5 Truth Table

Inputs										Outputs			
$K_9$	$K_8$	$K_7$	$K_6$	$K_5$	$K_4$	$K_3$	$K_2$	$K_1$	$K_0$	$Y_3$	$Y_2$	$Y_1$	$Y_0$
0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	1	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0	0	0	0	1	1
0	0	0	0	0	1	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0	0	0	1	0	1
0	0	0	1	0	0	0	0	0	0	0	1	1	0
0	0	1	0	0	0	0	0	0	0	0	1	1	1
0	1	0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	1	0	0	1	0	0	1

$K_n$  ( $n = 0, 1, \dots, 9$ ) represents the  $n$  keys. When  $K_i = 1$ , key  $i$  is depressed. It will be assumed that no more than one key is depressed simultaneously so that in any row every input except the  $i$ th key is a 0. From the truth table (8.5) we see that

$$Y_0 = 1 \text{ if } K_1 = 1 \text{ or } K_3 = 1, \text{ or } K_5 = 1 \text{ or } K_7 = 1 \text{ or } K_9 = 1$$

In Boolean notation

$$Y_0 = K_1 + K_3 + K_5 + K_7 + K_9.$$

Similarly,

$$Y_1 = K_2 + K_3 + K_6 + K_7$$

$$Y_2 = K_4 + K_5 + K_6 + K_7$$

$$Y_3 = K_8 + K_9. \quad \dots(8.6)$$

The above OR logic are implemented by diodes in Fig. 8.8(b). For example if key  $K_4$  is pressed the diodes connecting this line and outputs  $Y_2$  and  $Y_3$  will be short circuited to +5V giving the output 0101 which in turn identifies the decimal number 5.

#### d. Binary-coded-decimal (BCD) system

There are many schemes or codes, for representing a decimal number in binary form the most common being the binary-coded decimal (BCD). We know that representing decimal numbers as binary numbers leads to a long string of digits. Though easily handled in a digital system, they are difficult for a human being to remember. For example, the year  $1987_{10}$  in decimal system is equal to  $11111000011_2$  in binary system. We have little idea of what  $11111000011_2$  means until it is converted to the decimal system  $1987_{10}$ .

With the BCD system, each digit is coded or represented by four binary digits. Among various BCD systems the most popular is 8421 code. The MSB (most significant bit) has a weight of 8, while the LSB (least significant bit) has 1. This code is shown in Table 8.6 for each of the 10 digits in the decimal system.

**Table 8.6** BCD Code

Decimal	8	4	2	1 BCD
	D	C	B	A
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1



Each decimal digit forms a 4-bit word in the BCD code. A number containing four decimal digits would then be represented by four 4-bit words. For example,

$$1987_{10} = 0001\ 1001\ 1000\ 0111$$

$$\text{Since } 1 = 0001_2$$

$$9 = 1001_2$$

$$8 = 1000_2$$

$$7 = 0111_2$$

Though this conversion takes more bits (16 here) than direct conversion (where there are 11 bits) this system, no doubt, is more convenient to remember and handle.

### e. Decoder

A decoder performs the opposite function of an encoder and is an important digital system to perform various functions. For example, we wish to decode a BCD (binary-coded-decimal) instruction representing one decimal digit, say 9. This operation may be carried out with a four input AND gate excited by the 4 BCD bits. For example, the output of the AND gate in Fig. 8.9 (a) is 1 if and only if the BCD inputs are  $D=1$ ,  $C=0$ ,  $B=0$ , and  $A=1$ . Since this code represents the decimal number 9, the output is labelled *line 9*. The logic circuit

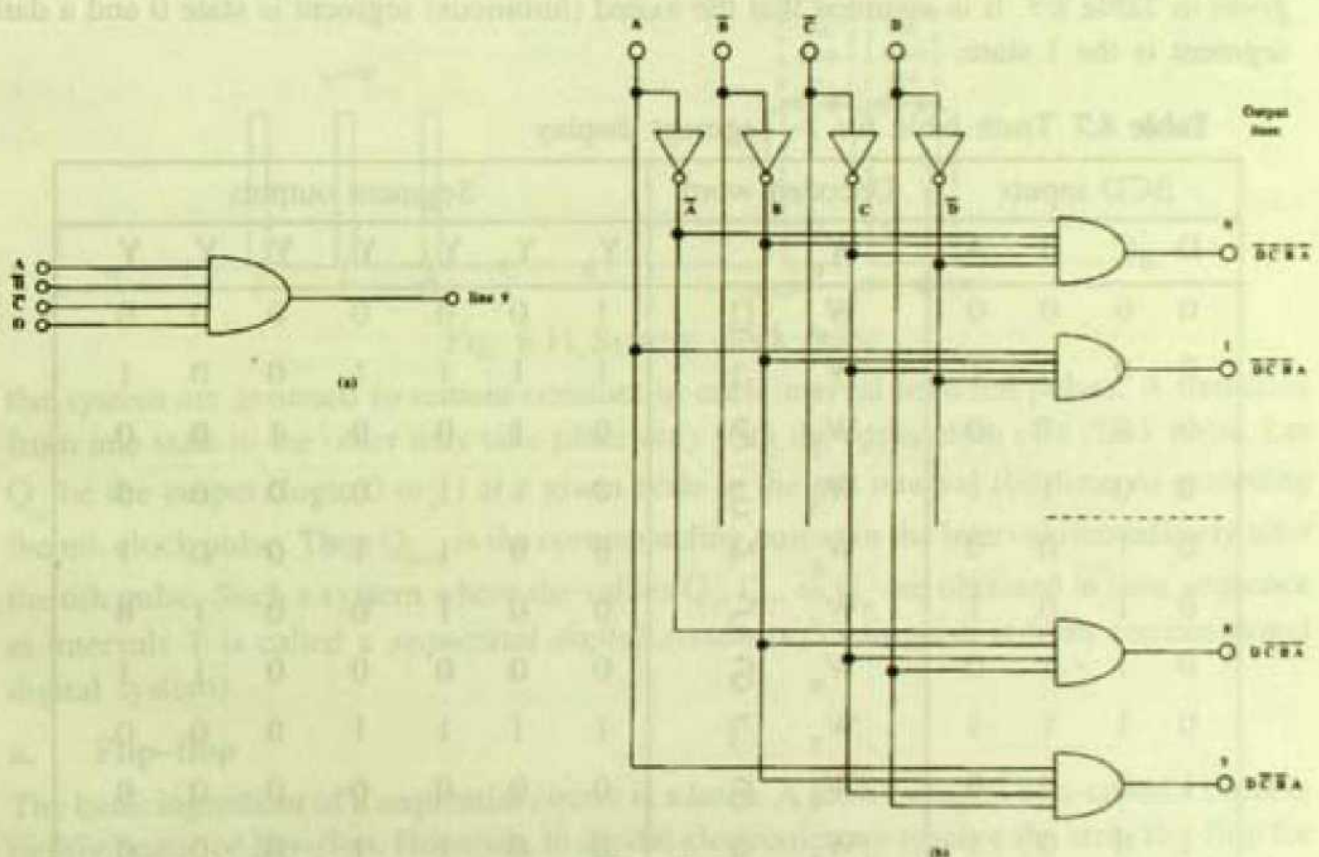


Fig. 8.9

(a) Output is 1 only when the input is  $D \bar{C} \bar{B} A$  i.e. 1001.

(b) a logic circuit for a BCD-to-decimal decoder.

for a BCD-to-decimal decoder is illustrated in Fig. 8.9 (b). This medium-scale-integrated (MSI) circuit has 4 inputs and 10 outputs and is called a *4-to-10-line decoder*. This means that a 4-bit input code selects 1 of the 10 output lines.

#### f. More about 7-segment Display

We shall now discuss the hardware that controls the operation of a 7-segment display (sec.7.3). A commercial IC, viz. 7447, is very useful for this purpose. This circuit has four input lines (DCBA) which represent a decimal digit in BCD. Seven of the output lines are used to drive a seven segment display that makes visible this decimal digit [Fig. 8.10]. The first

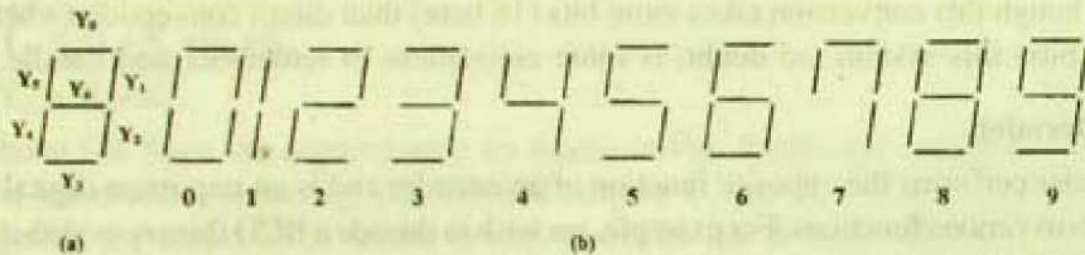


Fig. 8.10 (a) Identification of the segments, (b) the display from 0 to 9.

10 displays are the numbers from 0 to 9. The other 6 displays may be chosen by the designer. For example can make E, F, H, P etc. The truth table for the seven-segment display is given in Table 8.7. It is assumed that the excited (luminous) segment is state 0 and a dark segment is the 1 state.

**Table 8.7** Truth table for 7-segment display

BCD inputs				Decoded word		Segment outputs						
D	C	B	A	$W_n$		$Y_6$	$Y_5$	$Y_4$	$Y_3$	$Y_2$	$Y_1$	$Y_0$
0	0	0	0	$W_0$	0	1	0	0	0	0	0	0
0	0	0	1	$W_1$	1	1	1	1	1	0	0	1
0	0	1	0	$W_2$	2	0	1	0	0	1	0	0
0	0	1	1	$W_3$	3	0	1	1	0	0	0	0
0	1	0	0	$W_4$	4	0	0	1	1	0	0	1
0	1	0	1	$W_5$	5	0	0	1	0	0	1	0
0	1	1	0	$W_6$	6	0	0	0	0	0	1	1
0	1	1	1	$W_7$	7	1	1	1	1	0	0	0
1	0	0	0	$W_8$	8	0	0	0	0	0	0	0
1	0	0	1	$W_9$	9	0	0	1	1	0	0	0

The truth table is verified as follows. For the word 0 we see from Fig. 8.10 that  $Y_6 = 1$  and all other Y's are 0. For the word 8 all Y's are 0, and so on. The ROM (Read



Only Memory) can now be programmed to satisfy this truth table. For example,  $Y_6$  is 1 when  $W_0$  or  $W_1$  or  $W_7$  is 1, i.e.

$$\begin{aligned} Y_6 &= W_0 + W_1 + W_7 \\ &= \overline{D} \overline{B} \overline{C} \overline{A} + \overline{D} \overline{C} \overline{B} A + \overline{D} B C A \\ &= \overline{D} \overline{B} \overline{C} + \overline{D} B C A \quad (\because A + \overline{A} = 1) \end{aligned}$$

In this way the wired logic for segments  $Y_0, \dots, Y_7$  can be constructed.

### 8.3 Sequential Digital Systems

Until now we have considered only logic circuits whose outputs depended on the inputs. Now we shall introduce the concept of memory, so that the output will depend on the present and past values of the input signals. Such a circuit is called *sequential* because its output depends on a sequence of inputs. When this sequence is defined, or synchronized, by a system clock, we say that the circuit is *synchronous*.

A system clock (abbreviated ck) is shown in Fig. 8.11. The pulse width  $t_p$  is assumed to be small compared with  $T$ , the pulse train period. The binary values at each node in

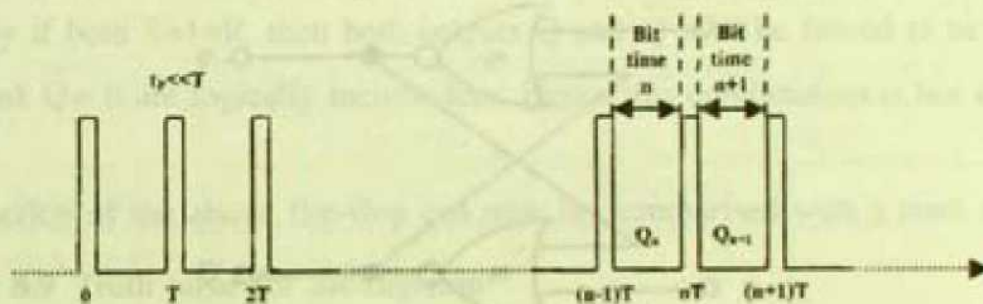


Fig. 8.11 System clock pulse

the system are assumed to remain constant in each interval between pulses. A transition from one state to the other may take place only with the application of a clock pulse. Let  $Q_n$  be the output (logic 0 to 1) at a given node in the  $n$ th interval (bit time  $n$ ) preceding the  $n$ th clock pulse. Then  $Q_{n+1}$  is the corresponding output in the interval immediately after the  $n$ th pulse. Such a system where the values  $Q_1, Q_2, \dots, Q_n$  are obtained in time sequence at intervals  $T$  is called a *sequential digital system* (to distinguish it from combinational digital system).

#### a. Flip-flop

The basic ingredient of a sequential circuit is a *latch*. A latch circuit is also called a bistable multivibrator or *flip-flop*. However, in digital electronics we reserve the term flip-flop for a more complex function that includes latch circuit in it. A latch has two stable states:— it either opens or closes. One of the stable states will be called SET or logic 1, whereas the other stable state will be called RESET, CLEAR, or logic 0. An input pulse is capable

of switching the output. The flip-flop (or simply FF) will remain stable in that state until another pulse switches the circuit back to the original state. Thus the name flip-flop signifies the ability of the circuit to change between two stable states. Since the state of an FF does not change following the removal of the input, the FF is essentially a *1-bit memory* or storage device.

We shall first examine a basic memory element, then we will extend this concept of the memory element to describe standard flip-flops.

### The NOR latch / SR Flip-flop

Fig. 8.12 shows two NOR gates connected to form a circuit called a *latch*. It is also known as set-reset or SR flip-flop. This circuit is used with the restraint that the input signals R and S are never 1 at the same time. The design engineer must ensure that the inputs satisfy the equation

$$R \times S = 0.$$

We immediately observe that there are two inverting amplifiers connected in a positive feedback configuration. In their high linear region this is an unstable circuit, with the result that one of the gates is driven to its high output condition and the other is driven to its low output condition; the output  $Q\bar{Q}$  becomes either 01 or 10. To understand this,

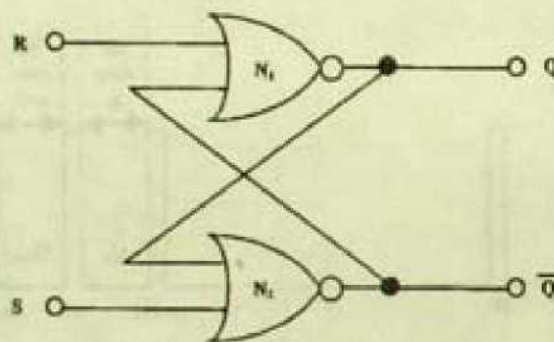


Fig. 8.12 A Latch with NOR Gates

**Table 8.8** Truth table for NOR

A	B	$Y = \overline{A + B}$
0	0	1
x	1	0
1	x	0

let the output Q be 1. As Q is the input to the NOR gate N2, according to the truth table for NOR gates, the output  $\bar{Q}$  must be in the 0 state. Thus if

$$Q = 1 \quad \text{then} \quad \bar{Q} = 0$$

There is another possibility. Instead of  $\bar{Q} = 0$ , let it be 1. As Q is the input to the



NOR gate N1, and because one of the inputs of this gate is 1, the output Q will be 0. Thus the other possible state of the circuit will be

$$Q = 0 \quad \text{if} \quad \bar{Q} = 1.$$

Thus this circuit possesses two stable states and hence called a bistable circuit.

### Action of S and R

Let us start with  $S=R=0$  and  $Q=1$ . From the truth table of NOR gate we notice that this is a stable condition in that no gate is in the process of changing its output; another way of saying this is that the output of each gate agrees with its inputs.

Now with  $Q=1$ ,  $\bar{Q}=0$  let  $R=1$  and  $S=0$ . In this case Q must change to 0 (don't care condition). Now the input to the gate N2 are  $Q=0$ ,  $S=0$ ; hence its output must change to 1.

Let us now change S to 1 and R to 0. With the same arguments as above we shall have  $\bar{Q}=0$  and now the inputs to N1 being 0,0, the output will be  $Q=1$ .

So far we have observed that the output of the above circuit flip-flops between its two stable states steered by the changes in the input.

Finally if both  $S=1=R$ , then both outputs Q and  $\bar{Q}$  will be forced to be in 0 state. But  $Q=0$  and  $\bar{Q}=0$  are logically inconsistent. Hence this combination is not allowed (or forbidden).

The action of the above flip-flop can now be summarised with a truth table (8.9).

**Table 8.9** Truth table for SR-flip-flop

Input		Output		Remark
S	R	Q	$\bar{Q}$	
0	0	Q	$\bar{Q}$	Unchanged
0	1	0	1	Reset
1	0	1	0	Set
1	1	?	?	not allowed

The SR FF can also be constructed by replacing NOR gates with NAND gates as shown in Fig. 8.13(a) with its truth table 8.10. Comparing this circuit with that of Fig. 8.12 and the corresponding truth tables we come to the conclusion that if S and R inputs of the circuit of Fig. 8.13(a) are inverted by two NOT gates, the function of the resulting circuit (Fig. 8.13 b) will be identical to that of the NOR SR-FF.

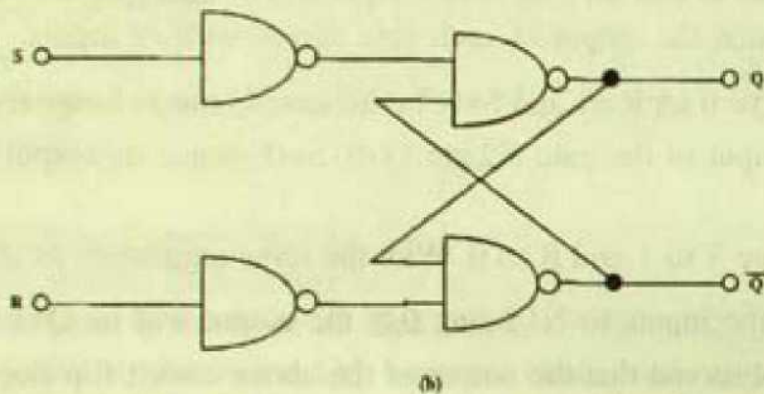
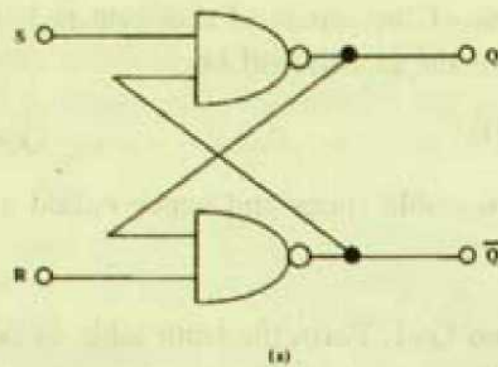


Fig. 8.13 (a) NAND SR-flip-flop (b) NOR equivalent NAND SR-FF

Table 8.10. Truth Table of NAND SR-FF

Inputs		Outputs		Remark
S	R	Q	$\bar{Q}$	
0	0	?	?	Not allowed
0	1	1	0	Set
1	0	0	1	Reset
1	1	Q	$\bar{Q}$	No Change

We finally sum up by saying that the FLIP-FLOP has two stable states. Hence it is called a *binary*, or *bistable*, multivibrator. Since it can store one bit of information (either  $Q=1$  or  $Q=0$ ), it is a *1-bit memory unit*, or a *1-bit storage cell*. Since this information is locked, or latched, in place, this circuit can also be called a *latch* as we said at the beginning.

### Clocked SR-FF

In sequential system it is very often necessary to set or reset the flip-flop in synchronism with clock pulses. This could be accomplished by a modified version known as clocked SR-FF and is shown in Fig. 8.14(a) together with its logic symbol in Fig. 8.14(b). This clock input decides whether the data is to be entered or to be ignored. To understand the



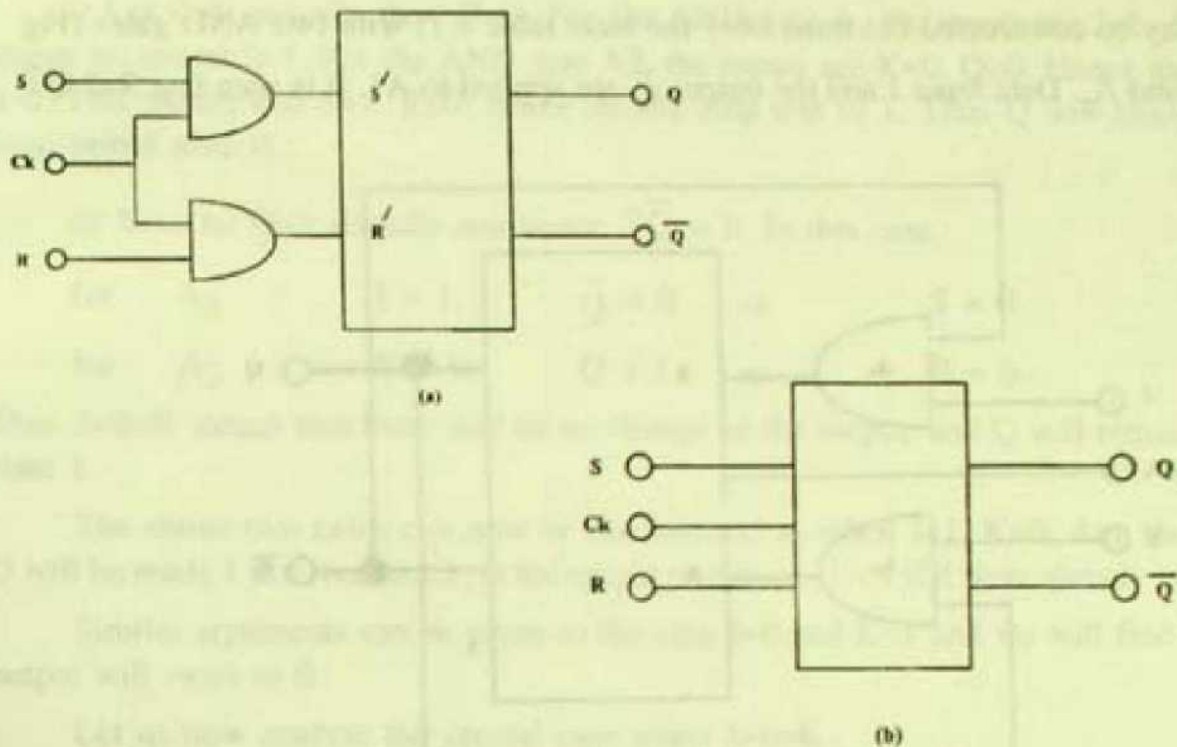


Fig. 8.14(a) clocked SR-FF; (b) circuit symbol

**Table 8.11** Truth table of clocked SR-FF

SR	S	R	Q
0	X	X	No Change
1	0	0	"
1	1	0	Set
1	0	1	Reset
1	1	1	Not allowed

action of the clock pulse we see that when the clock is low (i.e. logic), then both the AND gates are *disabled*. This means that outputs of both AND gates will be low making  $S'=0=R'$  and hence  $Q$  and  $\bar{Q}$  will change. Thus as long as  $ck=0$ , the flip-flop will not change its state and there would be no effect of  $S$  and  $R$  inputs on the outputs  $Q$  and  $\bar{Q}$ . Now, as the clock goes high, i.e.  $ck=1$ , then the AND gates are *enabled*. Depending on the states of  $S$  and  $R$ , the FF can *set*, *reset*, or there can be no change in the state. The resulting truth table is given in Fig. 8.11.

### b. J-K Flip-Flop

The major disadvantage of SR-FF is that the indeterminate condition occurs when both  $S$  and  $R$  inputs are simultaneously high. This difficulty is removed by a modified SR-FF known as JK-FF. This building block is obtained by augmenting the SR-FF (in whatever

way it may be constructed but must obey the truth table 8.7) with two AND gates (Fig. 8.15)  $A_1$  and  $A_2$ . Data input  $J$  and the output  $\bar{Q}$  are applied to  $A_1$ . It is seen that  $S = J \cdot \bar{Q}$ .

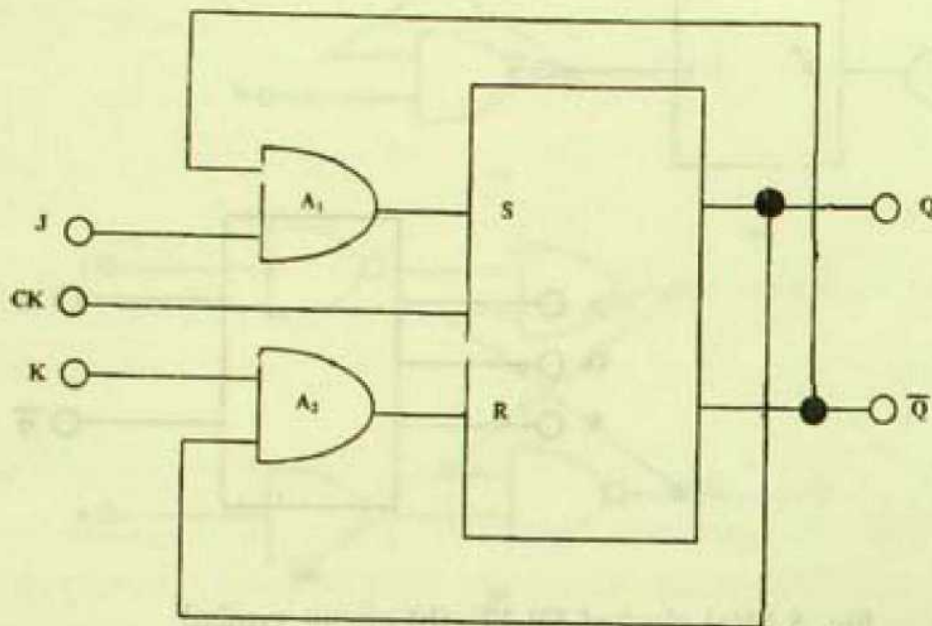


Fig. 8.15 JK-FF

Table 8.12 Truth table for JK FF

CR	J	K	Q	$\bar{Q}$	Remarks
0	X	X	Q	$\bar{Q}$	No Change
1	0	0	Q	$\bar{Q}$	"
1	1	0	1	0	Set
1	0	1	0	1	Reset
1	1	1	$\bar{Q}$	Q	toggle

Similarly  $R = K \cdot Q$ . Thus there are four possible combinations for the two data inputs  $J$  and  $K$ .

When the clock pulse is low the SR-FF is disabled and does not care for the inputs  $S$  and  $R$ . We shall consider the case only for  $ck=1$ .

(a) Let  $J=0=K$ . In this case both the AND gates are disabled and both  $S$  and  $R$  are low. The result is that there will be no change in the output, or in other words  $Q$  and  $\bar{Q}$  will remain in their earlier states.

(b) Let  $J=1, K=0$ . There are now two possibilities — i)  $Q$  is initially low and ii)  $Q$  is initially high.



i) Let  $Q=0$  initially, then  $\bar{Q}=1$ . For the AND gate  $A_1$ , the inputs are  $J=1$ ,  $\bar{Q}=1$ . Hence its output is 1. For the AND gate  $A_2$ , the inputs are  $K=0$ ,  $Q=0$ . Hence its output is 0. This means that  $S=1$ ,  $R=0$ ; hence the flip-flop sets to 1. Thus  $Q$  will change to 1 from initial state 0.

ii) Now let  $Q=1$  initially and hence  $\bar{Q}=0$ . In this case,

$$\text{for } A_1 : J = 1, \bar{Q} = 0 \Rightarrow S = 0$$

$$\text{for } A_2 : K = 0, Q = 1 \Rightarrow R = 0$$

Thus  $S=0=R$  means that there will be no change in the output, and  $Q$  will remain in the state 1.

The above two cases can now be summarized as when  $J=1$ ,  $K=0$ , then the output  $Q$  will be made 1 if it were not 1, or the output will remain 1 if it were already in state 1.

Similar arguments can be given to the case  $J=0$  and  $K=1$  and we will find that the output will *reset* to 0.

Let us now analyse the crucial case when  $J=1=K$ .

i) Let initially  $Q = 0$  and  $\bar{Q}=1$ :

$$\text{for AND gate } A_1 : J = 1 = \bar{Q} \Rightarrow S = 1$$

$$\text{for AND gate } A_2 : K = 1, Q = 0 \Rightarrow R = 0$$

Hence,  $Q$  should change to 1 (i.e. set condition).

ii) Let initially  $Q = 1$ , and  $\bar{Q} = 0$ :

$$\text{for } A_1 : J = 1, \bar{Q} = 0 \Rightarrow S = 0$$

$$\text{for } A_2 : K = 1, Q = 1 \Rightarrow R = 1$$

Hence  $Q$  should change to 0 (i.e. reset condition).

Thus whenever  $J=1=K$ , the output  $Q$  will change to 1 if it were initially at 0 and vice versa. This type of change in state is called *toggling* and the FF is said to toggle.

All the discussions so far made on JK-FF is summarized in the truth table (8.12).

### Preset and clear

It is really not necessary to use the AND gate  $A_1$  and  $A_2$  of Fig. 8.15, since the same operation can be performed by adding an extra input terminal to each NAND gate used instead of AND gates in a NAND SR-FF (Fig. 8.16a). The outputs  $Q$  and  $\bar{Q}$  are connected directly to the corresponding inputs as a positive feedback.

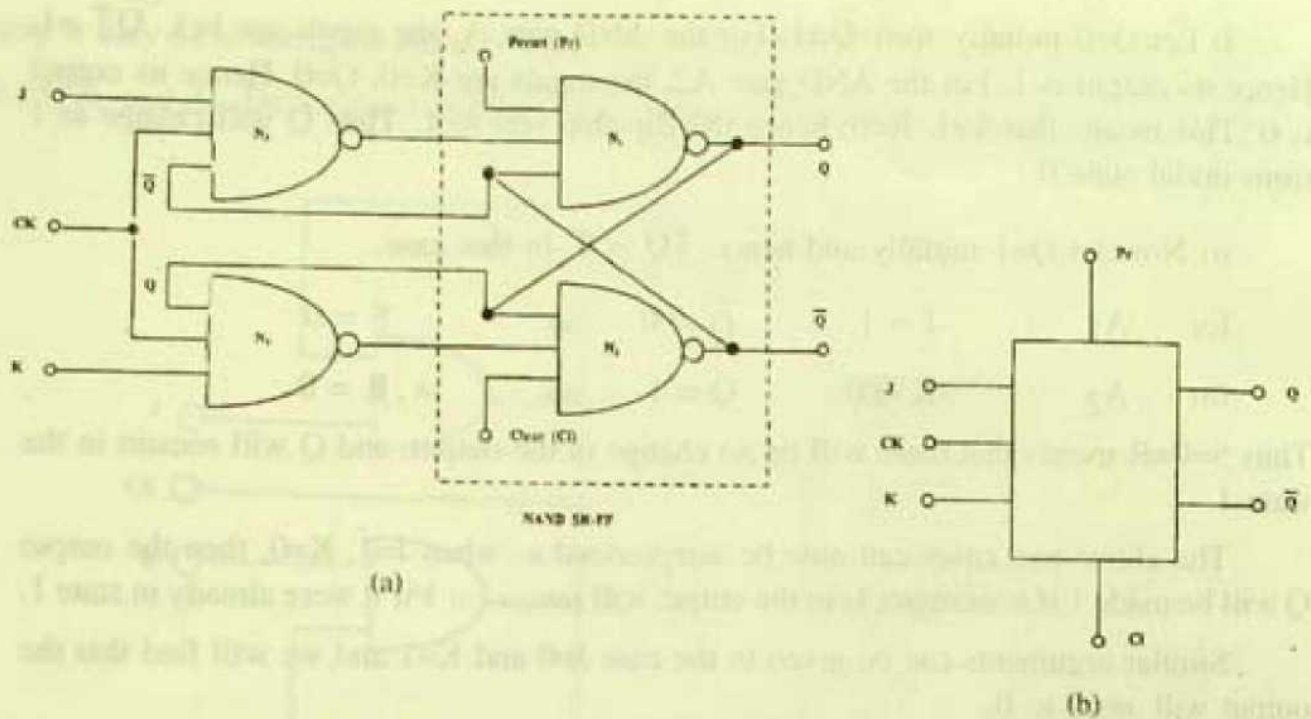


Fig. 8.16 JK-FF : (a) with preset and clear, (b) logic symbol.

So far we have discussed that if the output  $Q$  were at 1 or 0 then the state of  $Q$  will change in accordance with the changes in the inputs ( $J$  and  $K$  or  $S$  and  $R$ , and  $ck$  pulse). But at times it becomes necessary to be definite about the state of  $Q$ , i.e. either the flip-flop output  $Q$  will be at 0, called *clear* or at 1, called *preset*. In other words, we need to assign the initial condition, for example, we need to specify  $Q=0$  when  $ck=0$ . The addition of dashed inputs in Fig. 8.16(a) allows the initial state of the FF to be assigned.

The clear operation may be accomplished by putting clear input to 0, and preset input to 1. Since  $cl = 0$ , the output of  $N_2$  is  $\bar{Q}=1$ . Again, as  $ck=0$ , the output of  $N_3$  is 1 and hence all inputs to  $N_1$  are 1 which lead to  $Q=0$ , as desired. With similar arguments we can prove that  $Q=1$  if  $Pr=0$ ,  $Cl=1$ , and  $ck=0$ .

The preset and clear informations are called direct or *asynchronous* inputs, i.e. they are not in synchronism with the clock, and may be applied at any time in between the clock pulses. Once that state of the FF (i.e.  $Q$ ) is established asynchronously, the direct inputs must be continued at  $Pr=1$ ,  $Cl=1$ , before the next pulse arrives in order to enable the FF.

### C. Flip-flop

A clocked flip-flop may be used to store a bit i.e. it will keep in record the value of the input at a clock pulse and will hold the data until the next clock pulse comes. It latches the data and is often called a latch. However, the most common name of this type of FF



is called a D-FF (or delay FF). This is constructed by simply connecting a NOT gate between the inputs J and K, or S and R as shown in Fig. 8.17 together with its symbol and the

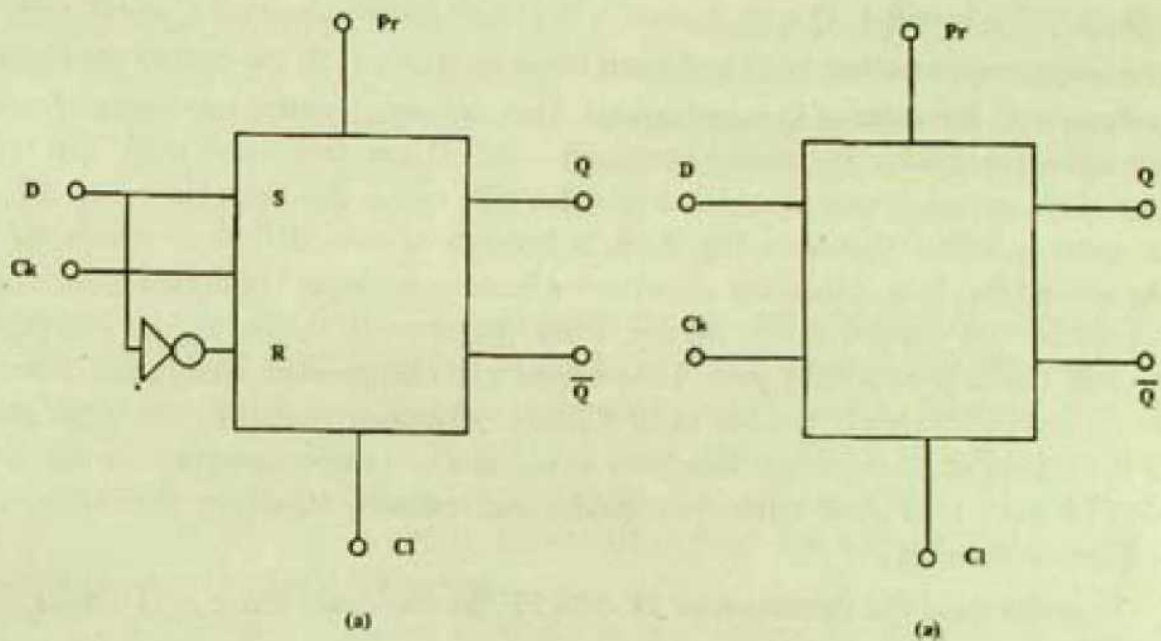


Fig. 8.17 D-FF: (a) circuit diagram, (b) symbol

**Table 8.13** Truth table for D-FF

Ck	$D_n$	$Q_{n+1}$
1	1	1
1	0	0
0	x	No change

the truth table 8.13. It should be noticed that there is no ambiguous state because  $S=1=R$  is not possible. It is seen from the truth table of SR-FF that  $Q_{n+1}=1$  when  $D_n=S_n=R_n=1$  and  $Q_{n+1}=0$  for  $D_n=S_n=R_n=0$ . Hence  $Q_{n+1}=D_n$ . The output  $Q_{n+1}$  after the pulse (bit time  $n+1$ ) equals the input  $D_n$  before the pulse (bit time  $n$ ), as shown in the truth table 8.13.

The D-type FF is a binary circuit used to provide delay. The bit (information) on the D line is transferred to the output at the next clock pulse, and hence this device functions as a 1-bit delay device.

#### d. Master-Slave Flip-Flop

In a clocked JK-FF, there is a restriction on the width of the clock pulse. The clock pulse must be smaller than the propagation delay  $\Delta t$  (approximate time interval between the application of an input pulse and the time when it reaches the output terminal) of the flip-flop i.e.  $t_p < \Delta t < T$ . Since the propagation delay  $\Delta t$  for IC FFs is very small ( $\sim 1$ ns) the

restriction  $t_p < \Delta t$  is not satisfied and the output becomes indeterminate. The reason behind this may be explained as follows. Let  $J=K=1$  and  $Q=0$ . If now a clock pulse is applied  $Q$  becomes 1 (toggling) after a time interval of  $\Delta t$ . At this moment  $J=K=Q=1$ , and since the clock pulse is still 1,  $Q$  will change to 0. Hence for the duration  $t_p$  of the clock pulse  $t_p$ , the output will oscillate back and forth between 0 and 1. At the end of the clock pulse i.e. when  $ck=0$ , the value of  $Q$  is *ambiguous*. This situation is called *race-around condition*. A circuit to avoid this race-around condition is called a *master-slave* (MS) flip-flop. The master slave principle can be utilized in either SR, or JK flip-flop. However, J-K MS is more popular and is shown in Fig. 8.18. It consists of two SR-FFs in which the output of the second flip-flop (called the *slave*) is fed back to the input of the first (called *master*). Clock pulses are applied to the master, while the slave is excited by an *inverted* clock pulse with the help of a NOT gate. Thus master can change state when  $ck=1$  whereas the slave can change its state only when  $ck=0$ . Clearly, the race-around difficulty is circumvented with the master slave topology. Needless to say this is a rather complex circuit; a typical JK-MS FF has a total of 40 transistors, diodes and resistors. However, with IC technology it is quite economical.

To understand the operation of JK-MS FF, let the clock pulse be 1 i.e.  $ck = 1$  and  $\overline{ck} = 0$ . Hence, the master is enabled and the slave disabled. Thus on the arrival of the clock pulse at the master input and with  $J = 1$ ,  $K = 0$  the master output is set to 1. At this instant the slave is disabled and the information on the master output (i.e.  $Q_M$ ) is

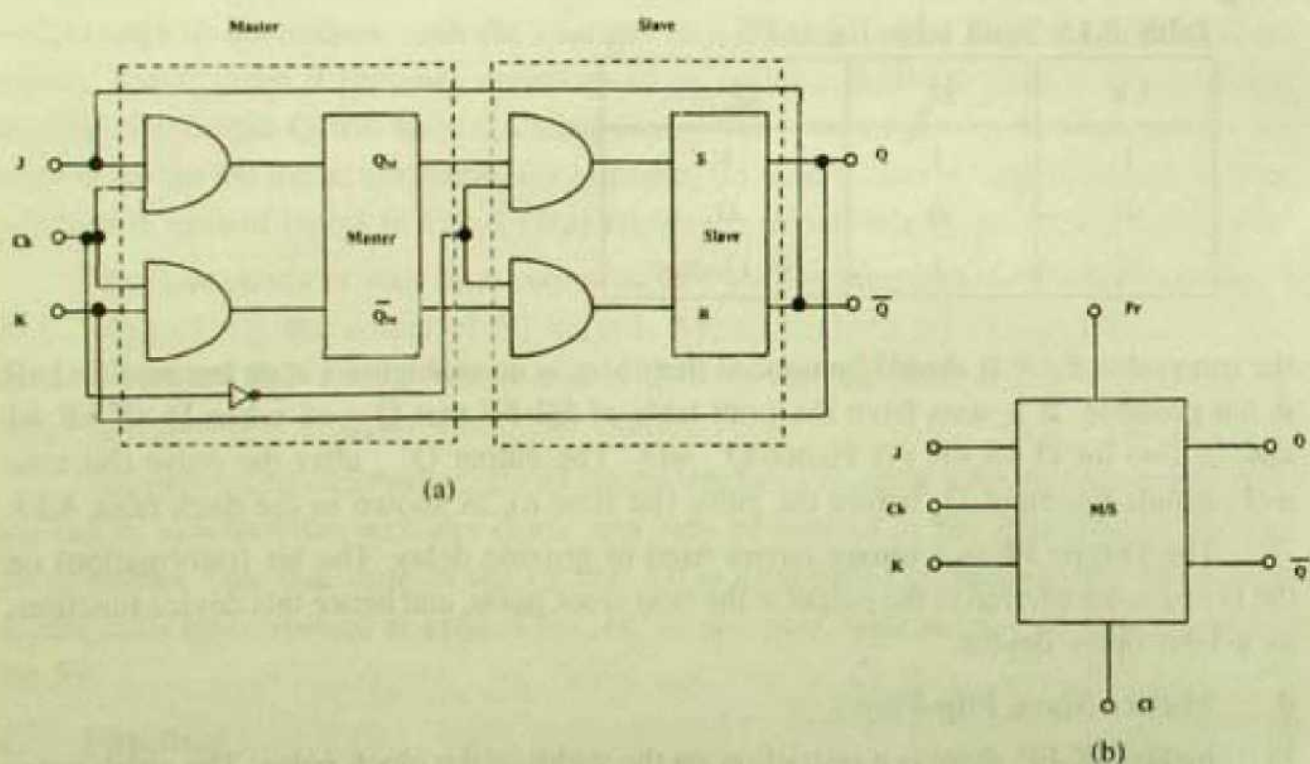


Fig. 8.18 JK- master slave flip-flop; (a) circuit diagram, (b) symbol  
stored in the slave inputs. It is brought to the output which is set to 1 ( $Q_M = 1 = S$ ,



$\overline{Q}_M = 0 (=R)$  only when the clock pulse goes low (i.e.  $ck = 0$ ). Similarly, for  $J = 0, K=1$ , the master is reset to 0 when  $ck = 1$  and the slave resets when  $ck=0$ . With  $J$  and  $K$  made 1, and the  $ck = 1$ , the output of the master toggles, i.e.  $Q_M \rightarrow \overline{Q}_M$ .  $Q_M$  being the input to  $S$  of slave FF,  $Q$  must follow  $Q_M$ . Hence, if  $Q_M$  toggles so will  $Q$ . Finally, with both  $J$  and  $K$  low there is no change of state in master latch and hence so is the case with slave latch.

To summarize, for a given JK input, the master responds as the clock goes high while the slave waits. But as the clock goes low, the information with the master is transferred to the slave. Thus the slave output follows the master output and there is no race around.

### e. Edge triggering

It is possible to have flip-flops (SR or JK) which respond on the leading or positive edge of the clock pulse. This type is often referred to as edge triggered, with the word *positive* left out. It is also equally possible to have negative edge triggering meaning thereby that the FF will respond at the falling edge of the clock pulse. Fig. 8.19 compares these two types of flip-flops, with a clock pulse waveform. The output for both JK- flip-flops are shown, together with circuit symbols. It is to be noticed that a small circle on the clock line indicates negative edge triggering.

The edge triggering circuit is normally used to reduce the duration of clock pulse. The method consists of a differentiating RC network. The RC circuit differentiates the

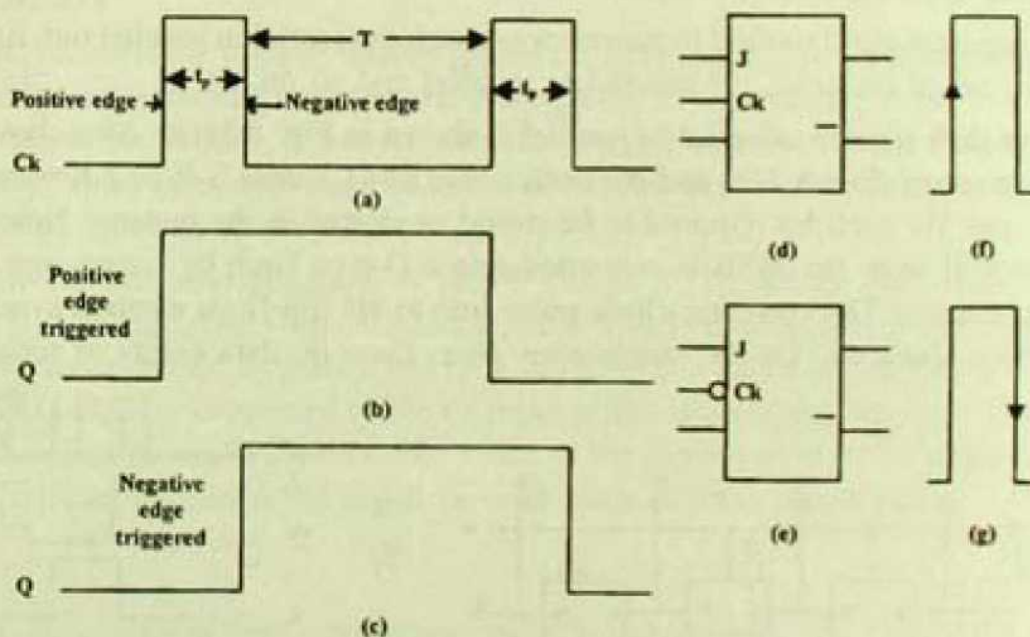


Fig. 8.19 Waveforms for triggering FF: (a) clock pulse (input), (b) Q (output) with positive edge trigger, (c) output with negative edge-trigger. Symbol of (d) positive edge-triggered JK-FF; (e) negative edge-triggered JK-FF, (f) positive clock pulse, (g) negative clock pulse.

rectangular clock pulse to produce positive and negative spikes (Fig. 8.20). The positive or negative spike may be conveniently taken by putting a diode, forward biased or reverse biased, as the case may be, across the output.

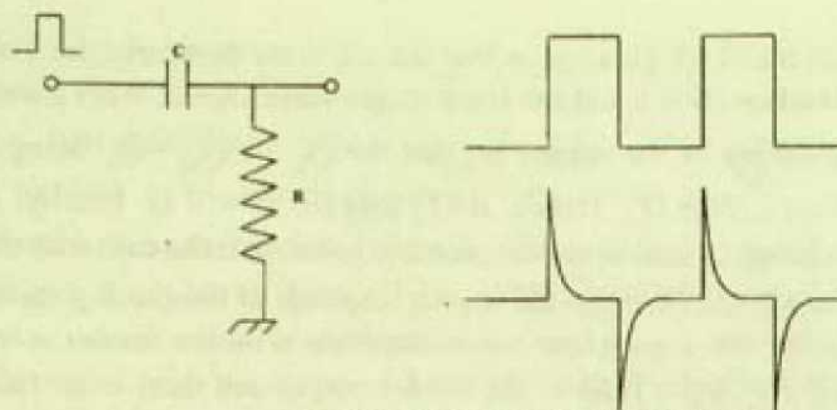


Fig. 8.20 Edge triggering circuit

## f. Registers

In performing binary arithmetic in a computer, it is usual to store the binary numbers as data words in registers, just the same way the kids put the digits in square books while they perform the arithmetic addition. The bit storage element is a simple flip-flop, so that we have as many flip-flops in the register as we have bits in the data word. However, means must be provided for shifting data bits into and out of the register. In a calculator a number is introduced by shifting the digits one by one. This shows that shift register must have memory in addition to the shifting action. The shifting is done by means of a pulse called the *command* signal, and a unit, consisting of a chain of flip-flops, used for storing or shifting the input data is called a *shift register*.

Shift registers are classified in various ways such as i) serial in parallel out, ii) parallel to serial, iii) serial to serial, iv) parallel to parallel and so on.

A 4-bit shift register of serial to parallel is shown in Fig. 8.21(a). Also shown is the wave form diagram (Fig. 8.21b) and the truth table (8.14). Either S-R or J-K master slave FF is used, one for each bit required to be stored or shifted in the register. Note that the stage which will store the MSB is converted into a D-type latch by connecting J and K through an inverter. The common clock pulse line to all flip-flops ensures synchronism of the shifting operation. On the single entry line, *Data in*, data enters in serial form.

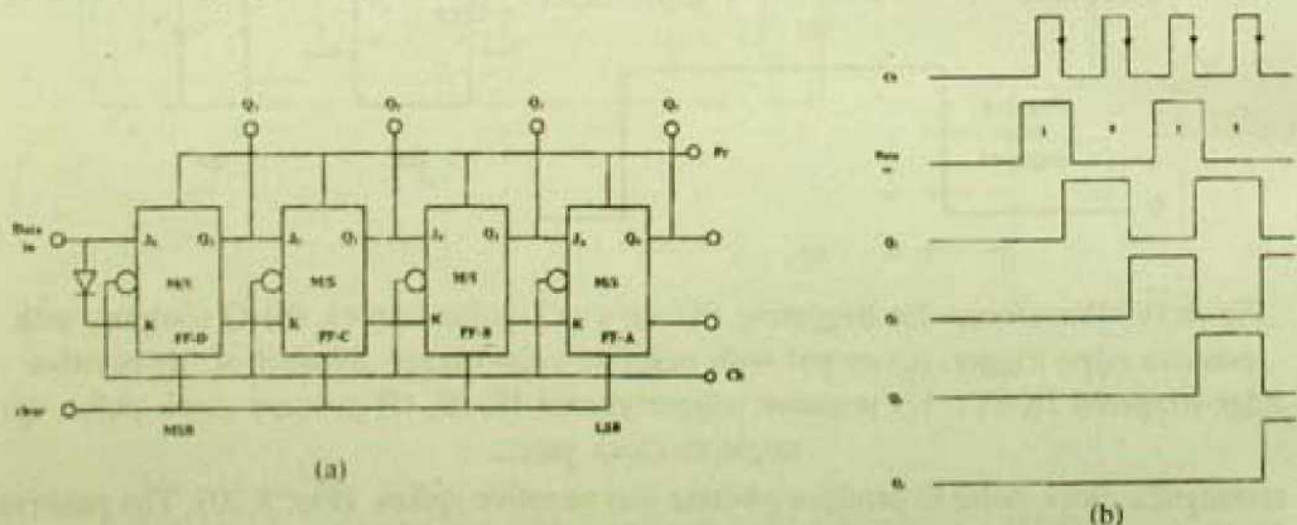


Fig. 8.21 4-bit serial-to-parallel shift register : (a) Block diagram (b) Timing waveforms



**Table 8.14** Truth table

ck	Serial Data in	$Q_3$	$Q_2$	$Q_1$	$Q_0$
1	1	0	0	0	0
0	2	1	0	0	0
1	3	0	1	0	0
0	4	1	0	1	0
		0	1	0	1

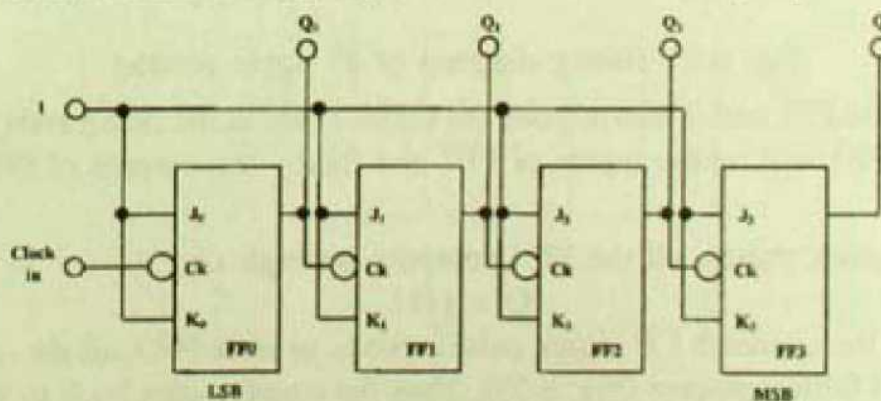
We shall now explain the operation of this simple 4-bit register by assuming that data 0101 is to be entered into it. The flip-flops are cleared by applying a clear input so that every output  $Q_0$ ,  $Q_1$ ,  $Q_2$  and  $Q_3$  is set to 0. Then we put  $Pr=1$ ,  $Cl=1$ . The LSB which is 1 here, is entered into FFD. This causes the master of FF-D to be set when clock is high; this is transferred to the slave and output when ck goes 0. The other flip-flops would have a 0.

At the second clock pulse state of  $Q_3$  is transferred to the master latch of FFC by the action of J-K FF. Simultaneously, the next bit (a 0 in the 0101 word) enters the master.

### g. Counters

Electronic counters are used in various digital applications. There are also many variations in the design of electronic counters. However, basic to all counters is the bistable circuit or flip-flop and the JK-M/S FF is most generally used because its output states are completely determined. Many of counters can be constructed, and any attempt to cover them all here would do an injustice to the subject.

From the characteristic table of JK-FF (Table. 8.12), we see that when  $J=K=1$ , the flip-flop will change state with each clock pulse. This is used to advantage in the simplest form of *binary counter*, as shown in Fig. 8.22. This is a chain of flip-flops, with  $J=K=1$ , and each  $Q$  output is connected to the ck input of the succeeding flip-flop. This is known as *asynchronous* or *ripple* counter. The effect of the input trigger pulse *rippling* down the counter is to set and reset the flip-flops with each alternate input pulse.



**Fig. 8.22** 4-bit binary ripple counter

To avoid problems of notation, the FFO flip-flop will always represent the LSB, and we shall proceed consecutively through the numerals to the MSB. It is the usual practice to draw logic block diagrams of counters with the input and LSB FF on the left and MSB on the right, but the binary word is written with MSB to the left and LSB to the right, viz.  $Q=Q_3Q_2Q_1Q_0$ .

The pulses to be counted are applied to the clock input of FFO. For all stages, J and K are connected to the supply voltage  $V_{cc}$  (logic 1 state) so that  $J=K=1$ . Since the FFs used are negative edge-triggered, the master will change from 1 to 0 and that the new state of the master is transferred to the slave when the clock rises from 0 to 1. When  $cl=0$ , all FFs reset and start from

$$Q = 0000.$$

When  $cl = 1$ , the counter is ready to go. Since FFO receives each clock pulse,  $Q_0$  toggles once per negative clock edge, as shown in the timing diagram of Fig. 8.23. At the falling end of the second clock pulse  $Q_0$  again toggles falling from 1 to 0 and so on. Now outputs

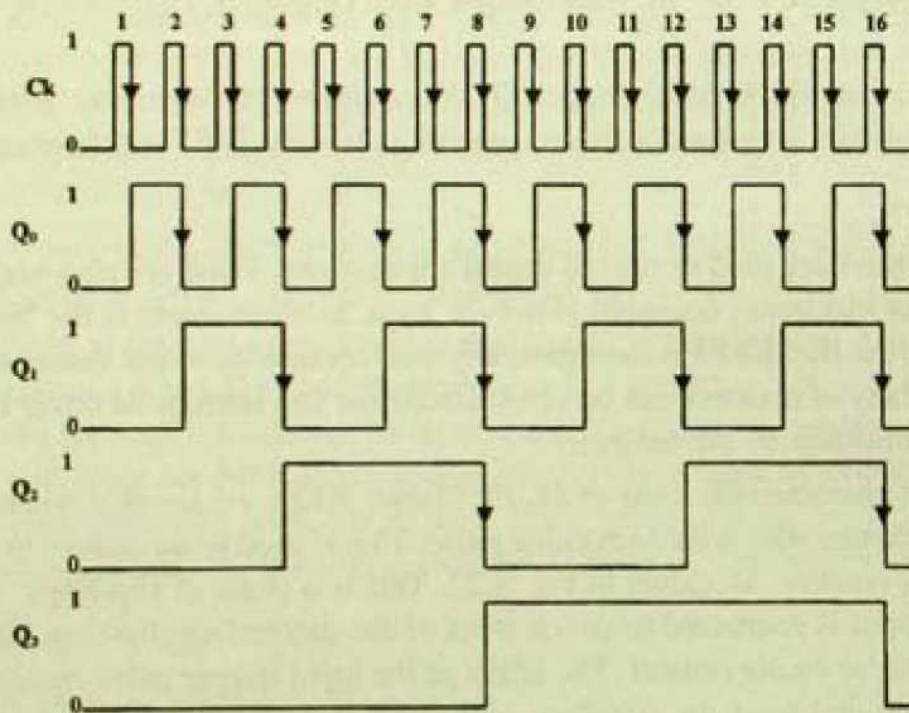


Fig. 8.23 Timing diagram of  $2^4$  ripple counter

of  $Q_0$  are inputs to FF1 and it will respond (or toggle) only at the rising ends of  $Q_0$ . Again, the outputs of FF1 will be the inputs of FF2 and finally the outputs of FF2 will be the inputs to FF3.

After 15 clock pulses, all the FF Q outputs are high i.e.

$$Q = 1111.$$

But when the sixteenth CP (clock pulse) arrives to reset FFO, all the others are also reset because of falling outputs (Fig. 8.23). Thus the counter goes back to its initial state and its can start recording fresh counts.



It is seen that a binary counter of  $2^4$  counts by  $(2^4-1)$ . Hence a chain of  $n$  binaries will count up to the number  $(2^n-1)$  and in the next CP it resets itself into its original state. Such a chain is referred to as a counter of *modulo*  $2^n$ .

### Down Counter

The counter that we described above is of the *count up* (also called up counter) type i.e. it starts counting from 0 upwards. However, it is sometimes required to initially store a number in a counter and then, as the pulses are applied to the counter, have it decrease the value stored in the counter and then give an indication when the counter is brought back to zero. This *countdown* type (also called DOWN counter) of counter is used in various games. It is also used in the control unit of a computer to indicate the end of a predetermined number of program steps.

A DOWN counter can be constructed from an UP counter by changing the trigger connection from the  $Q$  to the  $\bar{Q}$  output of each FF and is shown in Fig. 8.24. However, the encoded output is taken from the  $Q$  output.

A number is initially stored in the counter by using the preset and clear inputs.

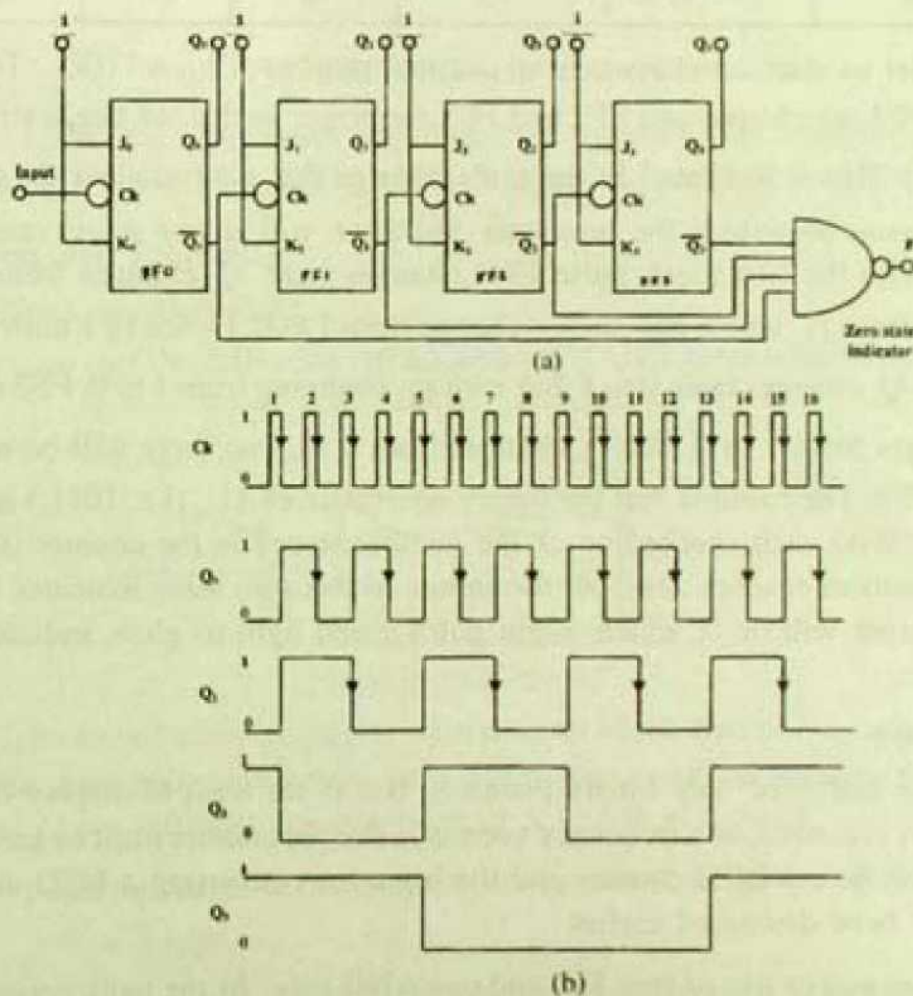


Fig. 8.24 4-bit Down counter (a) block diagram (b) timing waveforms

**Table 8.15** Truth table

State no.	Decimal no.	$Q_3$	$Q_2$	$Q_1$	$Q_0$
0	12	1	1	0	0
1	11	1	0	1	1
2	10	1	0	1	0
3	9	1	0	0	1
4	8	1	0	0	0
5	7	0	1	1	1
6	6	0	1	1	0
7	5	0	1	0	1
8	4	0	1	0	0
9	3	0	1	0	0
10	2	0	0	1	0
11	1	0	0	0	1
12	0	0	0	0	0

For example, let us start countdown from decimal number  $12_{10} \equiv 1100_2$ . To start with, FFs FFO and FF1 are cleared and FF2 and FF3 are preset so that we begin with  $Q = 1100$  ( $\equiv Q_3 Q_2 Q_1 Q_0$ ). This is indicated in the truth table as the *state number 0* i.e. the initial state of the counter because F, the zero state indicator, will be one if any one of its input is in state 0. With the first clock pulse FFO changes state,  $Q_0$  changes from 0 to 1, but input to FF1 is from  $\overline{Q_0}$ , which has made a change from 1 to 0. Hence FF1 must also change its state. Again  $Q_1$  changes from 0 to 1, but with  $\overline{Q_2}$  changing from 1 to 0, FF2 also changes state.  $Q_2$  changes from 1 to 0, but  $\overline{Q_3}$  changes from 0 to 1 so there will be no change at the output of FF3. The result is that the binary equivalent of  $11_{10}$  (i.e.  $1011_2$ ) is now stored in the counter. With each succeeding ck the number stored in the counter is reduced by 1. When the number reaches zero, all the inputs to the zero state indicator NAND gate is 1 and its output will be 0, which might put a green light to glow, indicating the end of the operation.

### Decade Counter

So far we have discussed only binary counters. But if we want to display the numbers, such as a digital voltmeter, or a frequency counter, a decade counter must be used. However, we shall discuss here a BCD counter and the logic for converting a BCD to its decimal equivalent has been discussed earlier.

The circuit makes use of four FFs and one AND gate. At the tenth count the outputs from the binaries must be  $Q = 1010$  and hence  $Q_3 = 1$ ,  $Q_2 = 0$ ,  $Q_1 = 1$ ,  $Q_0 = 0$ . But we want





- 8-5. Draw a logic diagram of a 4-to-10 line decoder using OR gates.
- 8-6. A truth table is written in the following form:

Inputs				Outputs			
$K_3$	$K_2$	$K_1$	$K_0$	$Y_3$	$Y_2$	$Y_1$	$Y_0$
0	0	0	1	0	1	1	1
0	0	1	0	1	1	0	0
0	1	0	0	1	1	0	1
1	0	0	0	0	0	1	0

Using diode matrix design an encoder that satisfies the above truth table.

- 8-7. Initial conditions for the counter shown in the figure is  $Q_0=1$ ,  $Q_1=0$ ,  $Q_2=0$ . Prepare a table of the readings  $Q_0$ ,  $Q_1$ ,  $Q_2$ ,  $J_2$  and  $K_2$  after each clock pulse. How many pulses are required before the system begins to operate as a divide-by-N counter? What is the value of N?

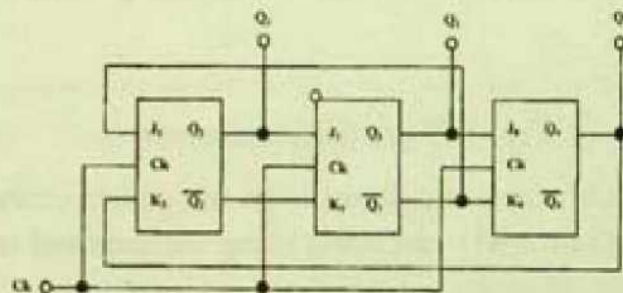


Figure for problem 8.7

- 8-8. Define an encoder. Indicate a diode matrix encoder to transform a decimal number into a binary code.
- 8-9. How will you augment a SRFF with two AND gates to form a JKFF? Give the truth table.
- 8-10. What is a truth table? Write down the output of the following truth table in terms of logic gates.

A	B	C	Y
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0



## Chapter 9

### Electronic Instruments

#### 9.1 Introduction

An instrument may be defined as a device for determining the value or magnitude of a natural (physical or chemical) quantity or variable. Measurement generally involves using an instrument as a physical means of determining a quantity or variable. The instrument serves as an extension of human faculties and in many cases enables a person to determine the value of an unknown quantity which his unaided human faculties could not measure.

The electronic instrument, as its name implies, is based on electrical and/or electronic principles for its measurement function. An electronic instrument may be a relatively uncomplicated device of simple construction such as a basic d. c. current meter. As technology expands, however, the demand for more elaborate, versatile and more accurate instruments increases and produces new developments in instrument design and application.

#### 9.2 Cathode ray oscilloscope

No electronics laboratory can run without a cathode ray oscilloscope (hereafter CRO). The television tube with which all of us are acquainted, is essentially a CRO. This instrument is used for measurement and analysis of waveforms and other phenomena in electronic circuits and provides a means of observing time-dependent voltages. CROs are essentially very fast x-y recorders that display an input signal versus another signal or versus time. The *stylus* of this plotter is a luminous spot that moves over the display area in response to input voltage. The CRO can present visual representations of many dynamic phenomena by means of transducers (a device that transforms a physical parameter into an equivalent electrical energy<sup>1</sup>). Recording of these happenings can be made by a special camera attached to the CRO for quantitative interpretation.

##### Basic CRO Operation

The major subunits of a general-purpose CRO are :

- i) Cathode ray tube, or simply CRT

---

<sup>1</sup> Strictly speaking a transducer is a device which, when actuated by energy in one transmission system, supplies energy in the same form or in another form to a second transmission system.

- ii) Vertical amplifier
- iii) Horizontal amplifier
- iv) Time base generator
- v) Trigger unit
- vi) Delay line
- vii) Power supply

The arrangement of these subunits are shown in Fig. 9.1 in block diagram form.

i) *Cathode ray tube, or CRT*

The CRT is, no doubt, the heart of the oscilloscope, and the rest only helps to operate the CRT. The essential components of a CRT are also shown in the Fig. 9.1 which are mounted inside a highly evacuated glass envelop. The CRT produces a sharply focussed beam of electrons and are accelerated to a very high velocity. This beam travels from its source (called the electron gun) and strikes the front of the CRT, inside of which is coated with a fluorescent

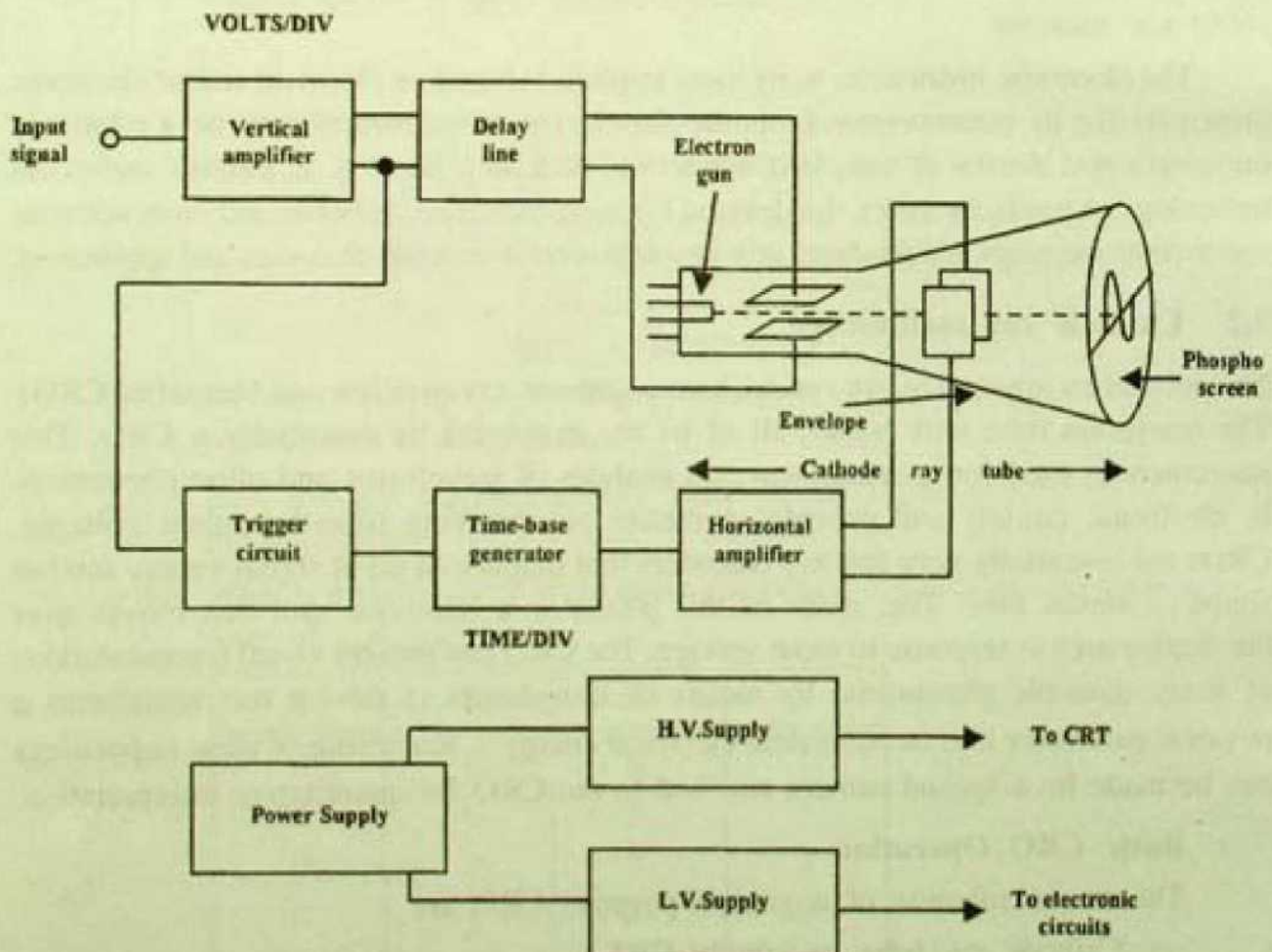


Fig. 9.1 Block diagram of a CRO



material (usually ZnS). The size of the beam spot depends on the electronic circuitry in the system.

The electron beam passes through a set of horizontal and vertical deflection plates. Appropriate voltages applied to the vertical and horizontal plates can control the position of electron spot anywhere on the screen.

### ii) Vertical amplifier

The signal waveform to be viewed on the CRT screen is applied to the vertical amplifier input. The gain of this amplifier is set by a calibrated input attenuator, usually marked as VOLTS/DIV. The output of a push-pull amplifier is fed to the vertical deflection plates of the CRT (Fig. 9.2), via a so-called *delay line*, with sufficient power to drive the CRT spot in the vertical direction.

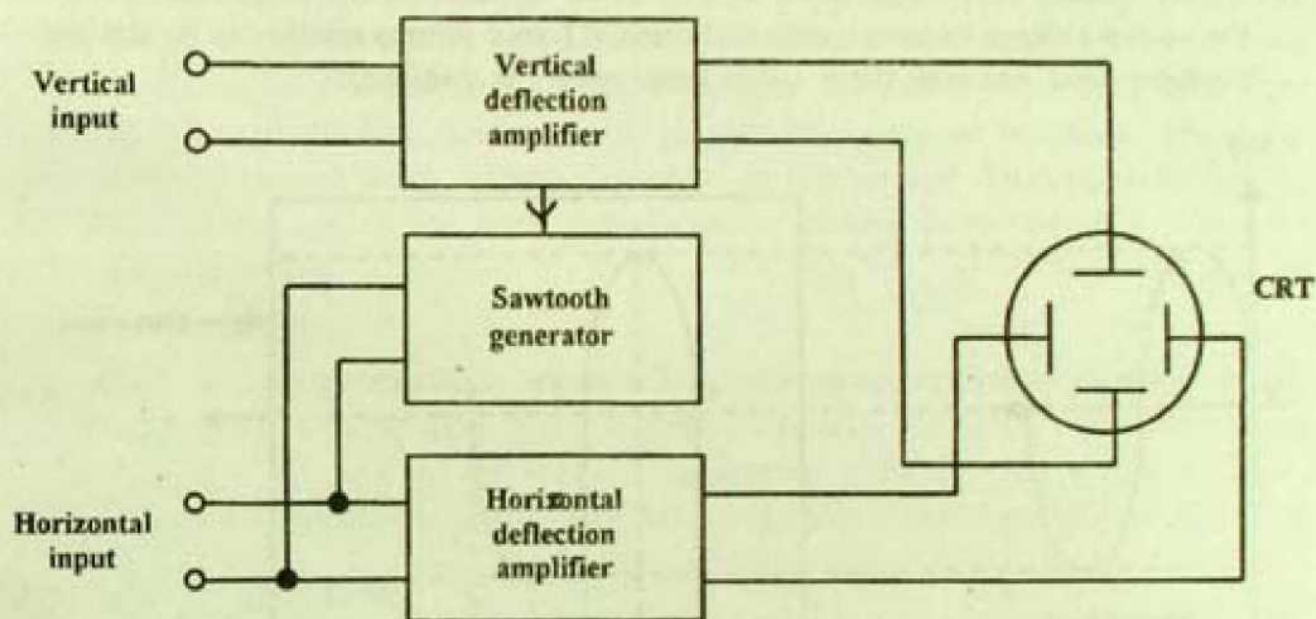


Fig. 9.2 Functional block diagram of CRO.

### iii) Horizontal amplifier

This amplifier comprises of a phase inverter. It produces two simultaneous output waveforms : a positive-going sawtooth (called run-up) and a negative going saw-tooth (called run-down). The positive-going sawtooth is applied to the right hand horizontal deflection plate and the negative going sawtooth to the opposite horizontal deflection plate. These voltages cause the beam of electrons to be swept across the CRT screen, along the positive x-axis, in time units that are controlled by the TIME / DIV control.

### iv) Time-base (TB) generator

The sweep generator, or more commonly known as time-base generator, develops a sawtooth waveform that is used as the horizontal deflection voltage of the

CRT. The positive going part of the saw-tooth waveform or sweep is linear, and its rate of rise is set by a front panel control marked as TIME / DIV. The saw-tooth voltage is fed to the horizontal amplifier. Of course, good linearity of the sweep is a major requirement for a reliable oscilloscope.

Simultaneous application of deflection voltages to both sets of plates causes the CRT spot to trace an image on the screen. This is shown in Fig. 9.3, where a sweep voltage is applied to the horizontal plates and, say, a sinewave signal is applied to the vertical plates. In this situation, the electron beam will be under the influence of two forces : one in the horizontal plane, moving the CRT spot across the screen at a linear rate, and the one in the vertical plane, moving the CRT spot up and down in accordance with the magnitude and polarity of the input signal. The resultant motion of the electron beam thus produces a CRT display of vertical input as a function of time. At the end of the sweep, the sweep voltage becomes zero and the CRT spot returns quickly to its starting position, and remains there till a new sweep is initiated.

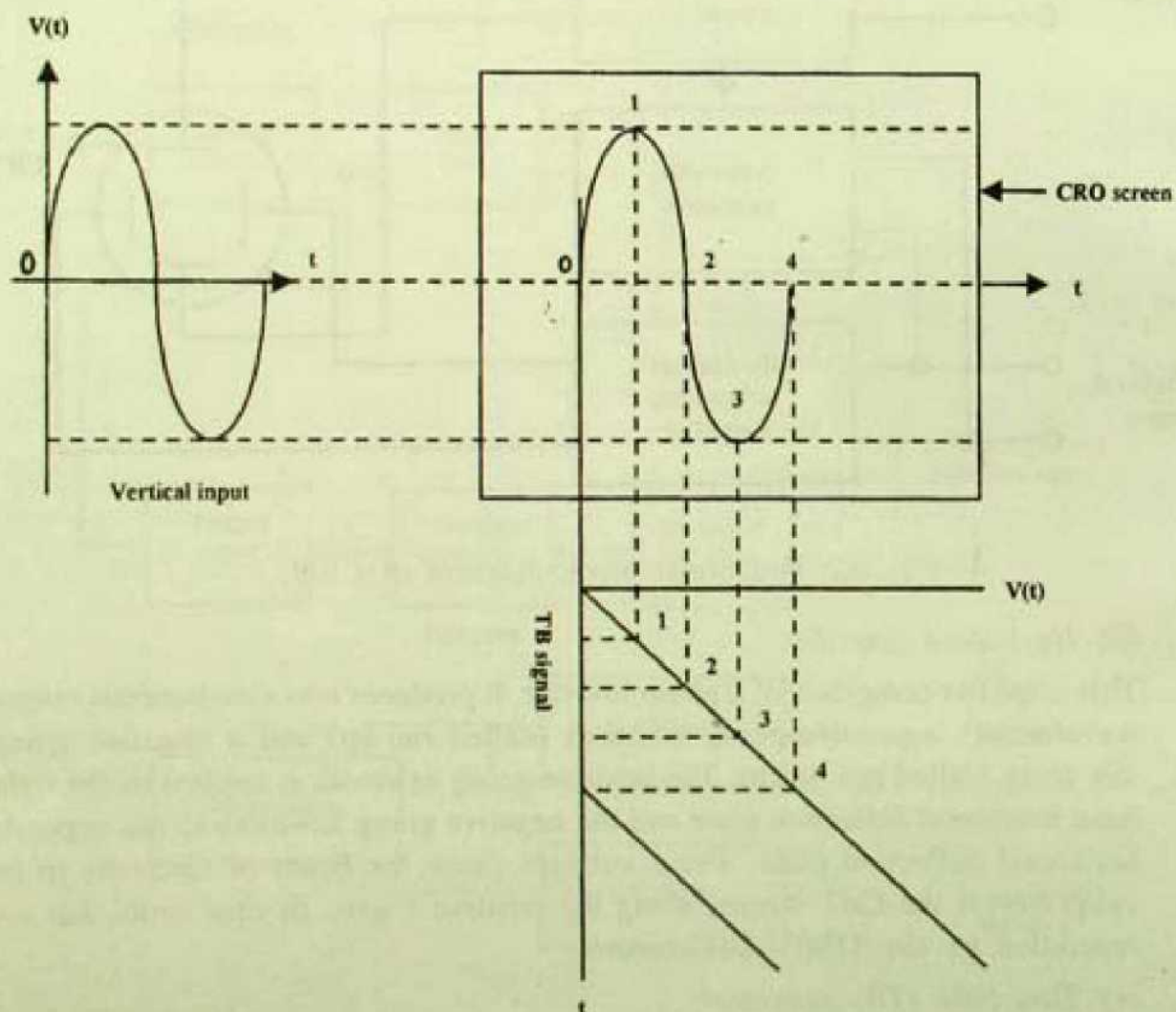


Fig. 9.3 Motion of a CRO spot



## Trigger Unit

If the input signal is of a repetitive nature, a stable CRT display can be maintained by starting each sweep at the same point on the signal waveform. To achieve this, a portion of the input signal is fed to a trigger circuit which produces a trigger pulse at some selected point on the input wave. This trigger pulse is used to start the time base generator, which in turn starts the sweep of the CRT spot from the left-hand side of the screen.

## Delay line

Normally, the leading edge of the input signal is used to activate the trigger generator which produces the trigger pulse and starts the sweep. This action takes place over a definite time interval (say 150 ns), so that the sweep is not initiated until after the leading edge of the input signal has passed. This then prevents the leading edge of the waveform to be displayed on the screen. A *delay line* is used to retard the arrival of the input signal at the vertical deflection plates until the trigger and time base circuits have had a chance to start the sweep of the beam. The delay line introduces a total delay  $\sim 250$  ns (variable) in the vertical deflection channel, so that the leading edge of the input signal can be viewed even though it was used to trigger the sweep.

## Power supply

There are two power supply sections. The high voltage section ( $\sim$ several Kilovolts depending on the size of the screen) is used to operate the CRT. The low voltage section produces dc voltages of various magnitudes, such as +5v,  $\pm 15$ v, +22volts etc. and is used to supply the voltages at various electronic circuits of the oscilloscope.

## 9.3 Digital Multimeter

A multimeter is an electronic instrument used to measure voltages, currents and resistances. There are two types of multimeters :— i) analog and ii) digital. In analog multimeters all the three measurements are done by the same galvanometer with suitable resistances connected with it in series or in parallel. A digital multimeter displays measurements as discrete numerals instead of a pointer deflection on a continuous scale as in analog devices. Digital readout is advantageous in many applications because it reduces human reading, parallax and interpolation errors, increases reading speed. Digital outputs are also suitable for recording and further processing in microprocessors and computers.

The basic device in a digital multimeter is a digital voltmeter (DVM). Optional features often include additional circuitry to measure current, resistance, inductance and capacitance. Other physical variables may be measured by using suitable transducers. Since the development and perfection of IC chips, the size, power requirements and cost of the DVM has been drastically reduced and become cheaper than its analog counter part.



Digital voltmeters may be classified into the following categories :—

- a) Ramp-type or single-slope
- b) Dual-slope
- c) Integrating
- d) Continuous-balance
- f) Successive approximation.

It should be mentioned that all these categories use somewhat complicated digital circuit to achieve its goal. However, we shall discuss only a simple but widely used one - the dual slope DVM.

### Dual-slope DVM

The principle of a dual-slope digital voltmeter is to convert the unknown voltage to a corresponding width (or interval) of a gating pulse. The gating pulse allows the high frequency (~MHz) clock pulses to pass through an AND gate and then they are recorded by a counter. Block diagram of such a dual-slope DVM is shown in Fig. 9.4(a).

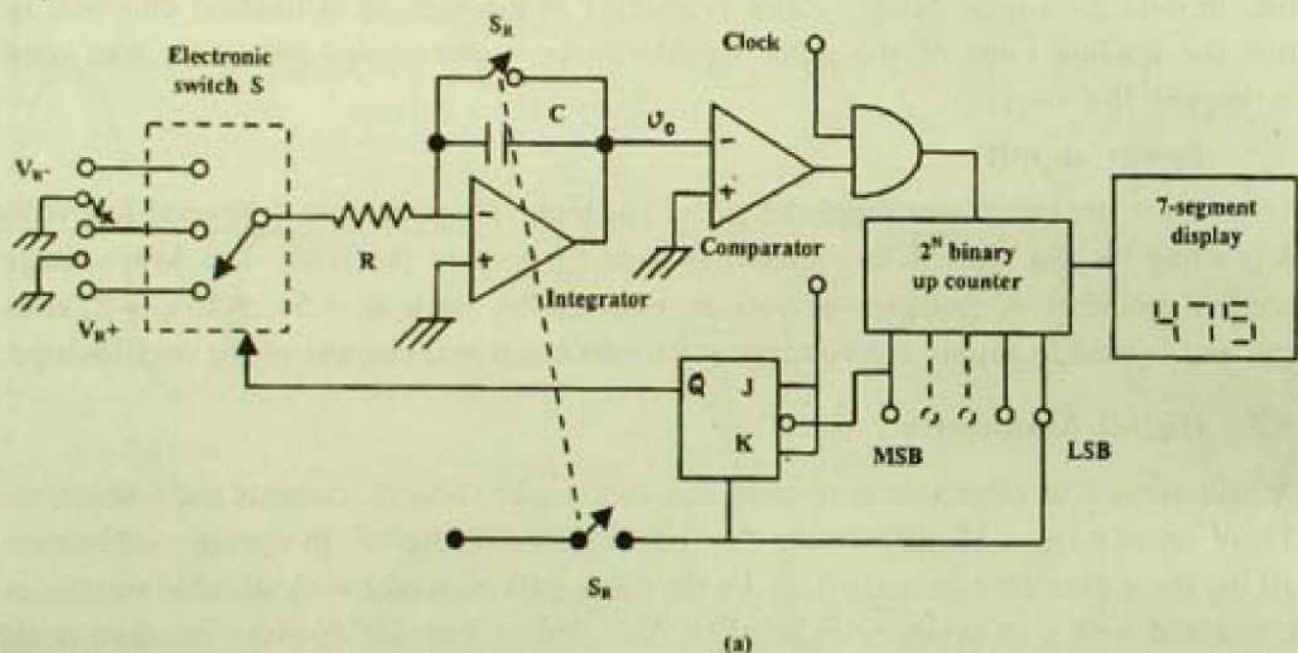


Fig. 9.4(a) Block diagram of a dual slope DVM.

Initially the electronic switch  $S$  grounds the integrator, and its output is set to zero and hence the output of the comparator is zero. Thus the AND gate is disabled. At this moment the JKFF (operated in the toggle mode) and the  $2^N$  binary counter is kept reset so that its output reads  $00 \dots 00$ . All these initial conditions are generally obtained by momentarily closing the start switch  $S_R$  at time  $t = t_0$ . At the same time the electronic switch  $S$  also connects to  $V_x$  to the input of the integrator (which generates a ramp voltage since  $V_x$  is constant) so that its output starts down with a slope proportional to  $V_x$  as shown in Fig. 9.4 (c). As the output



of the integrator crosses zero volt, output of the comparator switches to positive maximum at  $t = t_0$ , i.e. it changes to logic 1 state as indicated in Fig. 9.4 (d). This enables the AND gate.

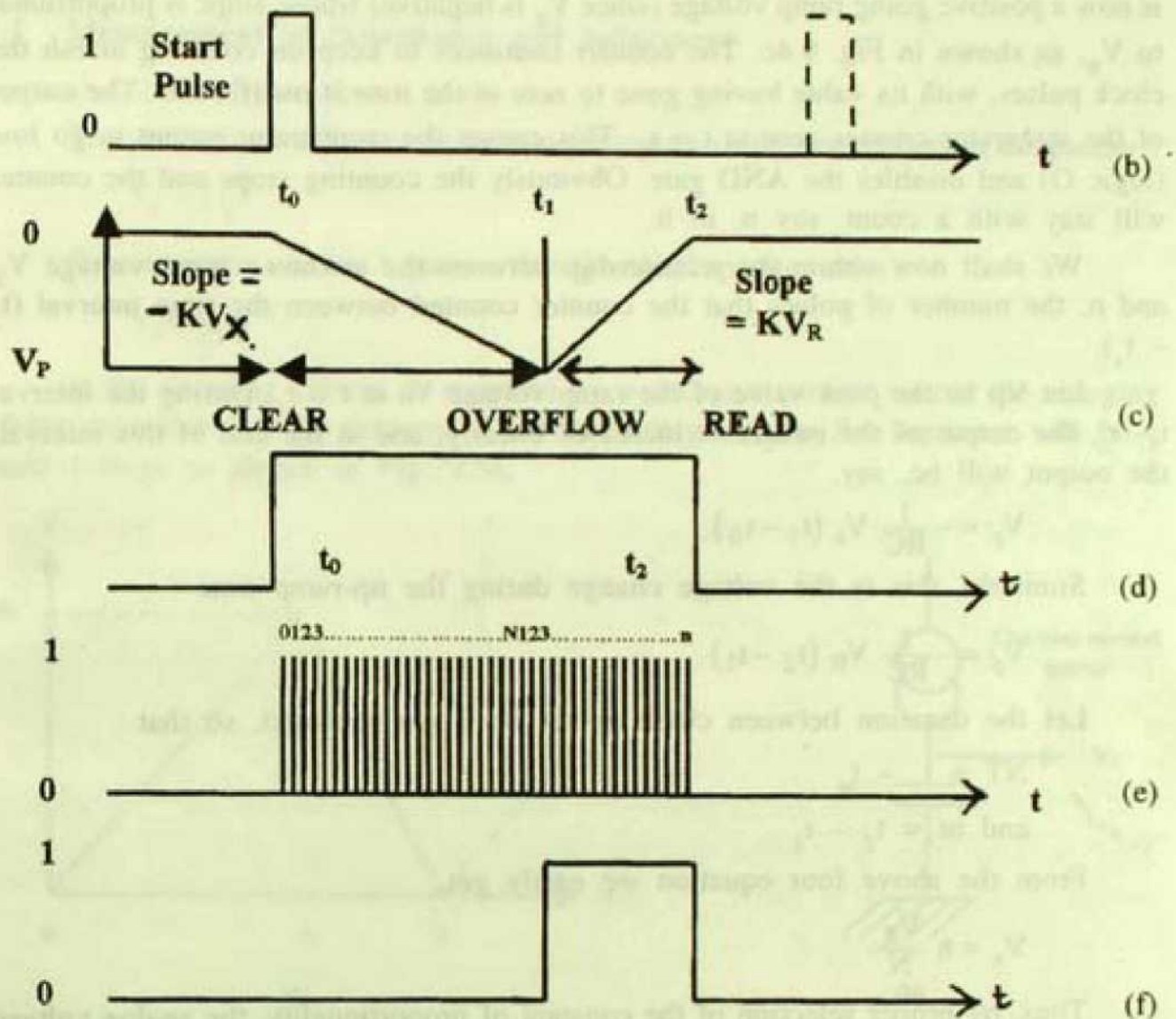


Fig. 9.4 (b) start pulse, (c) output of ramp generator (integrator), (d) output of voltage comparator, (e) gated clock pulses, (f) output of JK flip-flop.

Clock pulses (Fig. 9.4e) now reach the counter and counting proceeds. This counting will continue until the counter fills up its capacity of  $N$  pulses i.e. 11 . . . 11. This means that all  $Q$ 's of the counter becomes 1. It should be noted that during this interval, the output of the integrator increases continuously (Fig. 9.4c).

At the next clock pulse (at  $t = t_1$ ), the counter *overflows* i.e. all  $Q$ 's will go low and the counter will read 00 . . . 00. Obviously the last flip-flop (the MSB) will change from 1 to 0 for the first time. This will cause a toggling of the JKFF (negative edge-triggered). At this instant the output of the integrator reaches some maximum negative value.

Now, change in the output of JKFF (Fig. 9.4f) instantly activates the electronic switch S which disconnects  $V(x)$  from the circuit and in its place connects the reference voltage  $V_R$  (which is of opposite sign to that of  $V_x$ ). The output of the ramp generator is now a positive going ramp voltage (since  $V_R$  is negative) whose slope is proportional to  $V_R$ , as shown in Fig. 9.4c. The counter continues to keep on counting afresh the clock pulses, with its value having gone to zero at the time it overflowed. The output of the integrator crosses zero at  $t = t_2$ . This causes the comparator output to go low (logic 0) and disables the AND gate. Obviously the counting stops and the counter will stay with a count, say  $n$ , in it.

We shall now obtain the relationship between the unknown input voltage  $V_x$  and  $n$ , the number of pulses that the counter counted between the time interval  $(t_2 - t_1)$ .

Let  $V_p$  be the peak value of the ramp voltage  $V_o$  at  $t = t_1$ . During the interval  $t_1 - t_0$ , the output of the integrator increases linearly, and at the end of this interval, the output will be, say,

$$V_p = -\frac{1}{RC} V_x (t_1 - t_0).$$

Similarly, this is the voltage change during the up-ramp time :

$$V_p = -\frac{1}{RC} V_R (t_2 - t_1)$$

Let the duration between clock pulses be  $T$  (in seconds), so that

$$NT = t_1 - t_0$$

$$\text{and } nt = t_2 - t_1$$

From the above four equation we easily get

$$V_x = n \frac{V_R}{N}$$

Thus, by proper selection of the constant of proportionality, the analog voltage can be made equivalent to the count recorded. It should be noted that the measured binary equivalent of analog voltage is independent of the integrator time constant  $RC$  and the frequency (or period  $T$ ) of the clock pulses.

The DVM that we have described meets all the basic goals. Furthermore, this DVM can be the basis of a version with improved performance by just adding more hardware. By adding more electronic switching and OP AMPS, we would get automatic polarity, i.e. a reading of the absolute value of the voltage irrespective of polarity and a polarity indicator.

For voltages greater than the overload value applied to our DVM, we would have to use a precise resistive attenuator in order to get an *on scale* reading. DVMs are available that automatically switch attenuators and display the result in either a decimal-point change, or the units of the display are indicated.



Various other functions are available in commercial DVMs. Many of these can be used to measure both dc and ac voltages and currents as well as dc resistance. Hence they are called *digital multimeters*.

## 9.4 Measurement of capacitance and inductance

### a) Digital Capacitance Meter

The principle used in this meter is to generate a voltage equivalent to a capacitor :

$$V_c = \int \frac{i dt}{C}$$

If a constant current  $i$  passes through the capacitor then

$$V_c = \frac{i}{C} t$$

If the charging current is supplied by a constant current source the charging of the capacitor will be uniform. Thus the voltage across the capacitor will be a ramp voltage as shown in Fig. 9.5a.

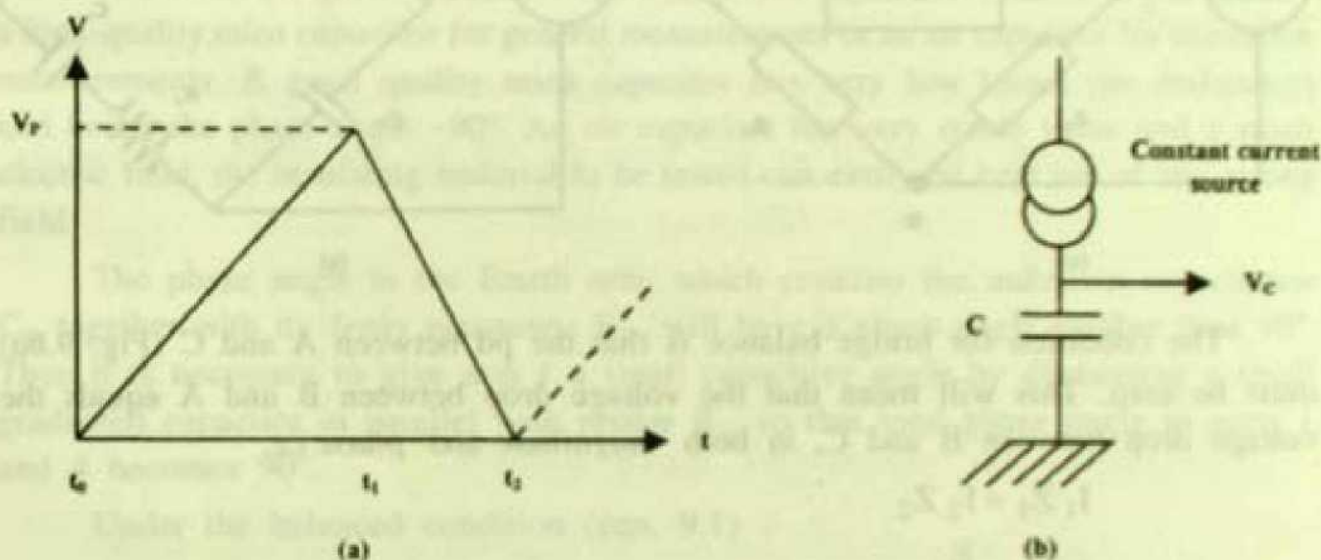


Fig. 9.5 (a) Ramp voltage across a capacitor, (b) charging the capacitor.

Charging of the capacitor continues till it reaches a saturation voltage  $V_F$  at a time  $t = t_1$ . At this moment, with the help of an electronic switch, the capacitor is made to discharge. The discharge continuous till the voltage across the capacitor goes to zero at a time  $t_2$ .

The measurement of this voltage can be done by a DVM and hence we can measure  $C$ . With similar principle we can also measure an inductance  $L$ .

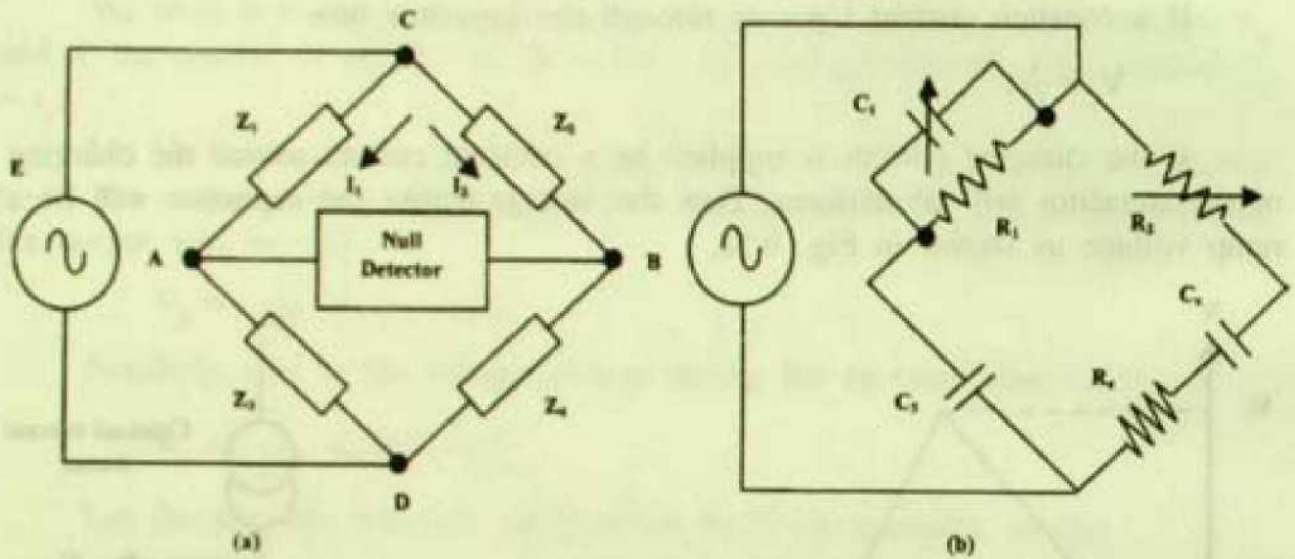
### b) Analog Capacitance Meter

In its basic form the measurement of an unknown capacitance or inductance is done by an ac bridge by comparing it with a known capacitance or inductance. There are various forms of ac bridge each having its own merits and demerits. In a commercial universal impedance bridge normally six forms of ac bridges are used.

### General form of ac bridge

The ac bridge is a natural outgrowth of dc wheatstone bridge and in its basic form consists of four bridge arms, a source of excitation, and a null detector. The power source supplies an ac voltage and the null detector is either a headphone or an ac amplifier.

The general form of ac bridge is shown in Fig. 9.6 (a) where  $Z_1$ ,  $Z_2$ ,  $Z_3$  and  $Z_4$  are the impedances. Balance condition is reached by varying one or more impedances of the bridge arms.



The condition for bridge balance is that the pd between A and C (Fig. 9.6a) must be zero. This will mean that the voltage drop between B and A equals the voltage drop between B and C, in both *magnitude* and *phase* i.e.

$$I_1 Z_1 = I_2 Z_2$$

$$\text{where } I_1 = \frac{E}{Z_1 + Z_3}$$

$$\text{and } I_2 = \frac{E}{Z_2 + Z_4}$$

Solving above three equations we get

$$Z_1 Z_4 = Z_2 Z_3 \quad \dots\dots(9.1)$$

or when using admittances instead of impedances

$$Y_1 Y_4 = Y_2 Y_3 \quad \dots\dots(9.2)$$

Equation (9.1) is the general equation of balance and is the most convenient form in most cases. It is wellknown that impedances are complex quantities in general, and if we write,

$$Z = |Z| \angle \theta$$



where  $|Z|$  is the magnitude and  $\theta$  the phase angle. In multiplication of complex numbers the magnitudes are multiplied while phase angles are added. Hence, eqn (9.1) can also be written as

$$|Z_1| |Z_4| \angle(\theta_1 + \theta_4) = |Z_2| |Z_3| \angle(\theta_2 + \theta_3) \quad \dots(9.3)$$

Eqn. (9.3) shows that two conditions must be met simultaneously when balancing an ac bridge :

- i) The products of the magnitudes of the opposite arms must be equal,
- ii) The sum of the phase angles of the opposite arms must be equal.

### Schering bridge

One of the most important ac bridges used extensively for the measurement of capacitors is the Schering bridge. This bridge is also useful for the measurement of phase angles (which is a property of an insulator).

The basic circuit is shown in Fig. 9.6(b). The standard capacitor  $C_3$  is usually a high-quality *mica capacitor* for general measurements or an air capacitor for insulation measurements. A good quality mica capacitor has very low losses (no resistance) and hence the phase angle  $\sim 90^\circ$ . An air capacitor has very stable value and a small electric field, the insulating material to be tested can easily be kept out of any strong field.

The phase angle in the fourth arm, which contains the unknown capacitance  $C_x$  together with its *leaky resistance*  $R_x$ , will have a phase angle smaller than  $90^\circ$ . Then it is necessary to give arm 1 a small capacitive angle by connecting a small graduated capacitor in parallel with resistor  $R_1$ , so that total phase angle in arms 1 and 4 becomes  $90^\circ$ .

Under the balanced condition (eqn. 9.1)

$$\frac{Z_x}{Y_1} = Z_2 Z_3$$

or 
$$R_x - \frac{j}{\omega C_x} = Z_2 Z_3 Y_1 = R_2 \left( \frac{-j}{\omega C_3} \right) \left( \frac{1}{R_1} + j\omega C_1 \right)$$

Equating real and imaginary terms, we find

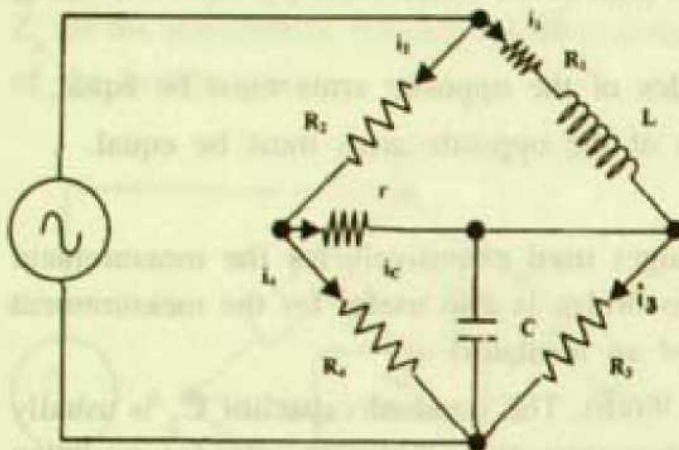
$$R_x = R_2 \frac{C_1}{C_3} \quad \dots(4)$$

$$C_x = C_3 \frac{R_1}{R_2} \quad \dots(5)$$

Thus the unknown capacitance can be determined.

### Anderson bridge

This bridge is one of the finest method of measuring inductance because it is applicable to the precise measurement of  $L$  over a wide range of values. This method requires a standard capacitor in terms of which the self inductance is expressed. The circuit diagram of an Anderson bridge is shown in Fig. 9.7.



$L$  = Self-inductance to be measured

$C$  = standard capacitor

$R_1$  = resistance including the resistance of  $L$ .

$r, R_2, R_3, R_4$  = Known non-inductive resistances.

Fig. 9.7 Anderson bridge for the determination of self-inductance.

Under balance conditions we find, from the Fig. 9.7, the following equations:

$$i_1 (R_1 + j\omega L) = i_2 R_2 + i_c r$$

$$i_3 R_3 = \frac{i_c}{j\omega C}$$

and  $i_c r + \frac{i_c}{j\omega C} = (i_2 - i_c) R_4, \because i_4 = i_2 - i_c$

Eliminating the currents  $i_1, i_2$  and  $i_c$  results

$$R_1 + j\omega L - j\omega C + \frac{R_2 R_3 r}{R_4} - \frac{R_2 R_3}{R_4} - j\omega C \cdot R_2 R_3 - j\omega C \cdot r R_3 = 0$$

Equating real and imaginary terms we find

$$\boxed{R_1 R_4 = R_2 R_3}$$

and

$$L = CR_3 \left( R_2 + r + \frac{R_2 r}{R_4} \right)$$

or

$$L = C [R_2 R_3 + r (R_3 + R_1)]$$

It should be noticed that it may be found impossible sometimes to obtain a balance by varying  $r$  and  $R_1$ . The formula for  $L$  with a value of  $R_2 R_3$  which satisfies the formula for  $R_1$  may require a *negative* value of  $r$ . It is thus desirable to know an approximate value of  $L$  and to take care that the product  $CR_2 R_3$  is less than  $\omega L$ . A balance is then possible with a positive value of  $r$ .



The impedances of the arms should be of the same order of magnitude (discussed under sensitivity of ac bridge). An approximate value of the unknown inductance should be estimated by this or some other method.  $R_2$  and  $R_3$  should then be chosen of the order of  $\omega L$ . If the experiment is then carried out an accurate value for  $L$  should result.

### Sensitivity of an ac bridge

In experiments with bridges we are interested in the conditions near those of balance, i.e. when  $Z_1 Z_4 = Z_2 Z_3$ . If a small deviation from the condition of balance results in a large change in the current through the null detector, the bridge is said to be sensitive.

Let the branches of a bridge network contain impedances as shown in Fig. 9.8.

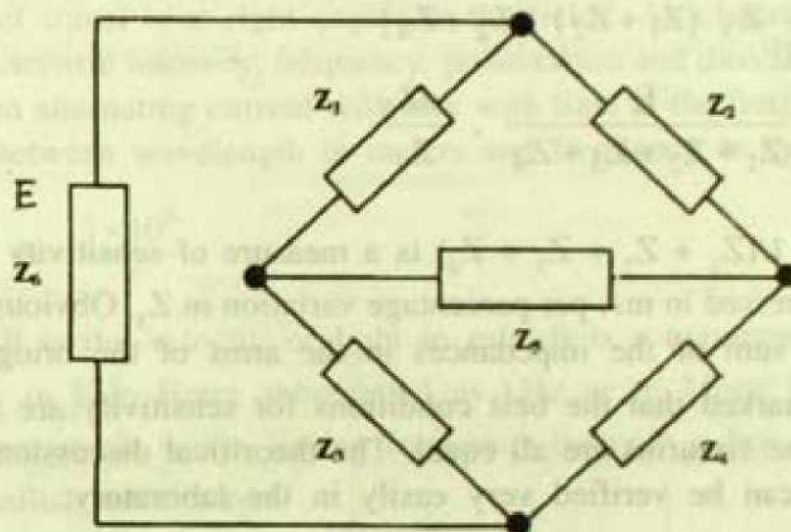


Fig. 9.8. Bridge Network

The current  $i$  in the detector, i.e. the arm containing  $Z_5$  is given by

$$i = \frac{(Z_1 Z_4 - Z_2 Z_3) E}{\Delta}$$

where  $E$  is the emf in the arm containing the impedance  $Z_6$ , which, for oscillators, may be quite high. The determinant  $\Delta$  is given by

$$\Delta = \begin{vmatrix} -Z_4 & -(Z_3 + Z_4) & Z_3 + Z_4 + Z_6 \\ Z_2 + Z_4 + Z_5 & Z_2 + Z_4 & -Z_4 \\ -Z_5 & Z_1 + Z_3 & -Z_3 \end{vmatrix}$$

This result is obtained by applying Kirchoff's laws to the network and requires the solution of simultaneous equations. The calculation is long but straightforward.

The bridge is balanced by varying, say,  $Z_3$ , i.e. by varying the resistance of one of the arms.

From equation for  $i$  we find

$$\frac{di}{dZ_3} = -\frac{Z_2 E}{\Delta} - \frac{Z_1 Z_4 - Z_2 Z_3}{\Delta^2} \cdot \frac{d\Delta}{dZ_3}$$

At the balance condition let  $\Delta = \Delta_0$ . Hence

$$\frac{di}{dZ_3} = -\frac{Z_2 E}{\Delta_0}$$

If we consider the  $Z_5$  and  $Z_6$  are negligible, which is generally true, then

$$\Delta_0 = Z_3 (Z_1 + Z_2) (Z_2 + Z_4)$$

and 
$$di = -\frac{E}{Z_1 + Z_2 + Z_3 + Z_4} \cdot \frac{dZ_3}{Z_3}$$

The factors  $= 1/(Z_1 + Z_2 + Z_3 + Z_4)$  is a measure of sensitivity of the bridge which is usually expressed in mA per percentage variation in  $Z_3$ . Obviously, sensitivity increases when the sum of the impedances in the arms of the bridge decreases.

It may be remarked that the best conditions for sensitivity are reached when the impedances in the six arms are all equal. The theoretical discussion of this point is complicated but can be verified very easily in the laboratory.



## Chapter 10

### Propagation of Electro-magnetic Wave

#### 10.1 Radio Waves

Electromagnetic waves travel through space with the velocity of light. These waves consist of electric and magnetic fields which are at right angles to each other and the direction of travel is at right angles to the fields. An electromagnetic wave will have its characteristic intensity, frequency, polarization and direction of travel. A wave produced by an alternating current will vary with time at the frequency of the current. The relation between wavelength in meters and frequency in cycles/sec is given by

$$\lambda = \frac{3 \times 10^8}{f} \quad (10.1)$$

where  $3 \times 10^8$  is the velocity of light in m/s. It is a common practice to express the frequency in Kilo Hertz abbreviated as kHz or in Mega Hertz abbreviated as MHz. Electromagnetic waves having frequency between a few kHz and 2000 MHz are usually called radio waves

The strength of a radio wave is expressed in terms of the voltage produced at a point by the electric field of the radio wave. It is usually expressed in microvolt per meter. The actual field strength produced at any point by an alternating sinusoidal wave varies sinusoidally from instant to instant and therefore the intensity of such a wave is generally expressed by its effective value which is 0.707 times the maximum value in a cycle. The strength of the wave measured in microvolts per meter (or  $\mu\text{V/m}$ ) in space is exactly the same voltage that the magnetic field of the wave induces in a conductor of one meter long by sweeping the conductor with the velocity of light. Sometimes the radio waves having signal strength as low as  $0.1 \mu\text{V}$  per m are highly useful, occasionally field strength of the order of  $1000 \mu\text{V}$  per m is required for satisfactory reception, but usually the useful signal strength lies in between these two extreme ends.

The direction of the electric field  $\vec{E}$  is called the direction of polarization of the wave. When the electric flux lines are vertical, the wave is called vertically polarized.



When the electric flux lines are horizontal and magnetic flux lines are vertical, the wave is called horizontally polarized. A plane which is parallel to the mutually perpendicular electric and magnetic flux is called the wavefront. The wave travels in the perpendicular direction on the wavefront but whether it moves forward or backward depends on the relative direction of the electric field  $\vec{E}$  and magnetic field  $\vec{B}$ . The direction of propagation is along  $\vec{E} \times \vec{B}$  vector. With the change of the direction of either of the electric or magnetic flux, the direction of propagation is reversed but reversing the direction of both electric and magnetic flux, the direction of propagation will not change. By attaching an antenna of appropriate size, the electromagnetic waves can be broadcast and received at a distance away. The radio, microwave, infrared and visible light portions of the electromagnetic spectrum are all used for transmitting information from one point to another by modulating amplitude, phase or frequency of the wave. The frequency band designated and listed by International Telecommunication Union is given in Figure 10.1.

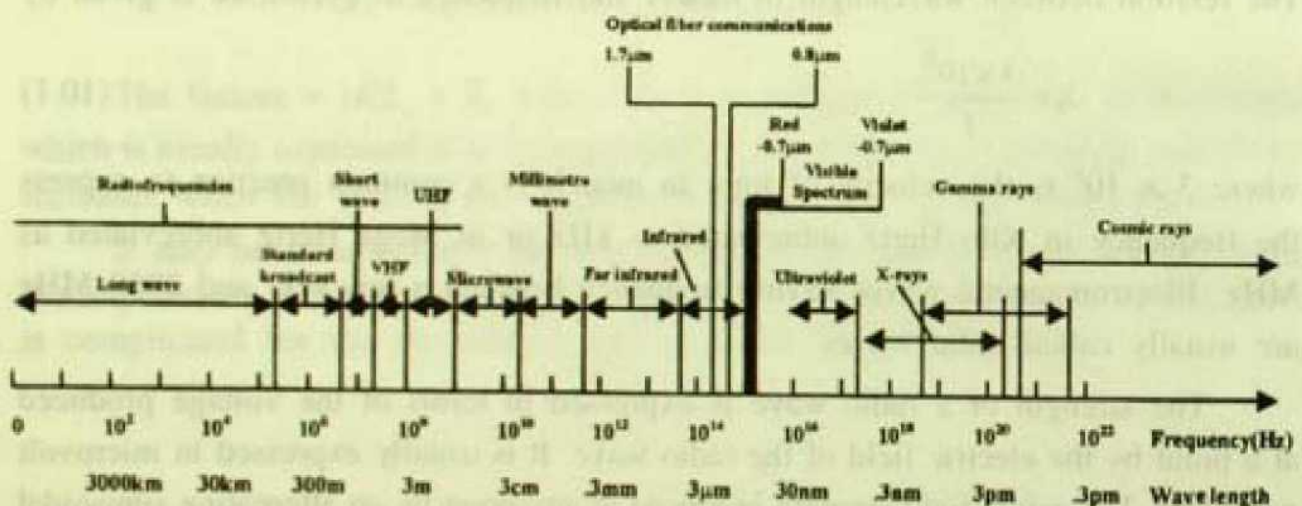


Fig. 10.1 Electromagnetic spectrum used for communication

### Classification of Radio Waves

All radio waves travel away from their point of origin. They become attenuated or weakened due to the spreading out of the wave. Moreover, the energy of the wave may be absorbed by the ground or by the ionized region in the upper atmosphere known as ionosphere. The waves may be reflected or refracted by the ionosphere or by the conditions at the lower atmosphere or by the ground. The resulting situation is quite complex and it differs widely at different frequency ranges of the radio wave. Radio waves at different bands have their distinctive uses. They are summarized in table 10.1.



**Table 10.1**
**CLASSIFICATION AND USES**

Class	Frequency range (Wave length range)	Typical use
Very low frequency (VLF)	10-30 kHz (30-10 km)	Long distance point to point communication
Low frequency (LF)	30-300 kHz (10-1 km)	Long distance point to point service, marine, navigation etc.
Medium frequency (MF)	300-3000 kHz (1000-100 m)	Broadcasting, marine communication, navigation etc.
High frequency (HF)	3-30 MHz (100-10 m)	Moderate and long distance communication of all types, broadcasting
Very high frequency (VHF)	30-300 MHz (10-1 m)	Short distance communication, television, frequency modulation, radar, navigation
Ultra high frequency (UHF)	300-3000 MHz (100-10 cm)	Short distance communication, radar, television etc.
Super high frequency (SHF)	3000-30,000 MHz (10-1 cm)	Radar, radio relay, navigation

Frequencies above 2000 MHz are generally called microwave frequencies

## 10.2 Earth and Atmosphere

### *Standard atmosphere*

The atmosphere is the gaseous region surrounding the earth. It contains gases and water vapour. It is a mixture of about 79% nitrogen, 20% oxygen and 1% of other gases and it contains water vapour varying from 0 to 5 per cent by volume. The temperature, pressure and water vapour content of the atmosphere decrease with increasing height from the ground surface when averaged over a long period of time. Normally the temperature decreases upward from the ground and reaches a turning height called the tropopause where the temperature stops decreasing. The temperature of the tropopause remain almost constant. The region from the ground surface upto the tropopause is called the troposphere and it extends upto about 15 km from the ground surface. The adjacent region above the tropopause is called the stratosphere. The radio wave propagation in the atmosphere is entirely influenced by the meteorological conditions of the atmosphere namely temperature, air pressure and water vapour.

The standard atmosphere for a region of a country is taken as a hypothetical atmosphere in which the properties are arbitrarily chosen to fit certain average conditions. Thus in a typical case (i) the tropopause is at a height 11500 meters. (ii) with increasing height from 0 to 16000 meters, the air pressure changes from about 760 mm to 87 mm. (iii) water vapour pressure changes from 7.7 mm to 0.01 mm. (iv) temperature changes from  $+15^{\circ}\text{C}$  to  $-55^{\circ}\text{C}$ . and (v) the variation of the refractive index  $\mu$  of the atmosphere is indicated by  $(\mu-1) \times 10^6$  as 324 to 42 only.

### *Electrical Properties of Earth*

The relative dielectric constant ( $\epsilon$ ) and electrical conductivity ( $\sigma$ ) of the earth are widely different at different places. Different kinds of soils, sea water, fresh water, forestation offer widely different values, as shown in table 10.2.

**Table 10.2**

### **TYPICAL GROUND CONSTANTS**

Types of Ground	Relative Dielectric Constant	Conductivity $\mu\text{S} / \text{cm}$
Sea water	81	45000
Fresh water	80	100
Rich soil	20	100
Forestation	13	50
Rocky/Sandy soil	10	20
Cities	5	10

In case of radio wave propagation, at medium and long waves, the capacitive reactance of the earth is much greater than the resistivity  $\frac{1}{\sigma}$  of the earth. Hence the earth may be considered to be purely resistive. But at a frequency of the order of 10MHz or more, the impedance offered by the ground is primarily capacitive. In determining the values of earth conductivity and dielectric constant for use in assessing the attenuation of a radio wave due to losses in the ground, suitably weighted average of these quantities should be used. There is a depth below the earth's surface to which there exist ground currents of appreciable magnitude. The depth of radio wave penetration depends upon the frequency, dielectric constant  $\epsilon$  and conductivity  $\sigma$ . It may be as low as a few feet at the short wave end and may be a few hundred feet at long waves. So, the ground constants are not materially affected by variations of conditions at the ground surface due to temporary rain fall etc.



### 10.3 Regions involved in Radio Wave Propagation

Bare earth surface helps in the propagation of a radio wave which is vertically polarized. This radio wave supported at its lower edge by the ground, is of practical importance for broadcasting at lower frequencies.

Well above the earth's surface from about 70 km to few hundred km, there exists the upper atmosphere which is stratified or ionized at different levels by external ionizing radiations. This ionized layers may help in the propagation of radio wave by bending the wave path. This ionized regions, commonly known as ionospheric layers, account for practically all very long distance radio communications.

Apart from the above two regions which help in radio wave propagation, there is the troposphere which extends upto 15 km from the earth's surface through which radio wave may propagate. Tropospheric propagation usually contains two components. One of these is a direct ray from the transmitter to the receiver, while the other is a ground reflected ray. Radio wave propagation at frequencies above about 30 MHz i.e. the frequencies used in television, frequency modulation, radar etc. is normally tropospheric propagation.

#### Classification of Radio Wave Propagation

Propagation of radio waves from the transmitter to the receiver may be in any one of the following two ways

- A. Ground wave propagation
- B. Sky wave propagation or Ionospheric propagation

Ground wave propagation is the propagation of a radio wave which results because of the presence of the ground. The ground wave may further be sub-divided into two categories —

- i) Surface wave
- ii) Space wave or tropospheric wave

#### *Surface Wave*

This wave travels along the surface of the earth. It is vertically polarized i.e. the electric vector of the electromagnetic wave is vertical. Such a propagation takes place when the transmitter and the receiver are both close to the ground surface. Surface wave propagation is of importance for medium and long wave signals. All medium wave signals during day time use surface wave propagation.

#### *Space Wave*

This wave travels from the transmitter to the receiver through the space i.e. earth's troposphere. The space wave is constituted by two components — one, the line of sight or direct wave and the other, the ground reflected wave as shown in Figure 10.2.



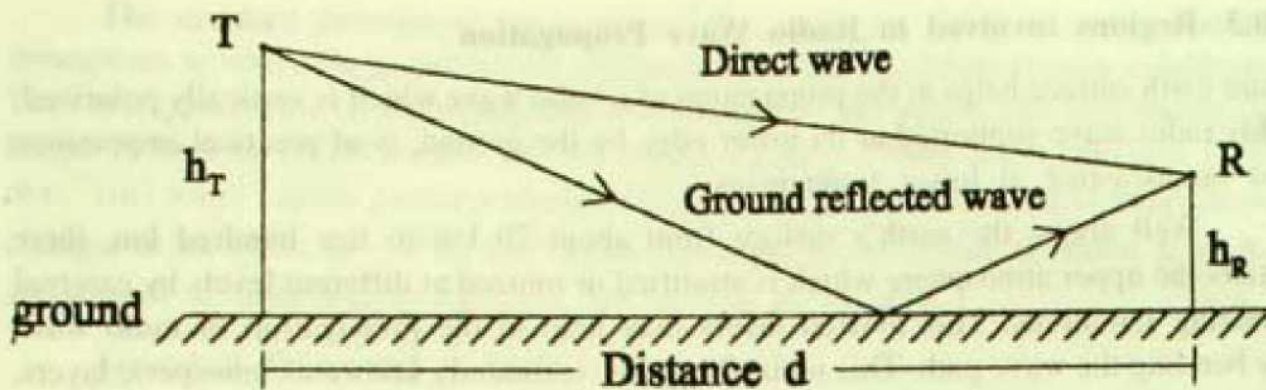


Fig. 10.2 Space wave propagation

The simplified diagram assumes a flat earth and neglects the curvature of the radio wave due to the variation of the refractive index of the earth's atmosphere with height from the earth's surface. The transmitter T is at a height  $h_T$  from the ground and similarly  $h_R$  is the receiver height from ground surface. Space wave propagation is normally used in television, frequency modulation broadcasting etc. at a frequency above 30 MHz.

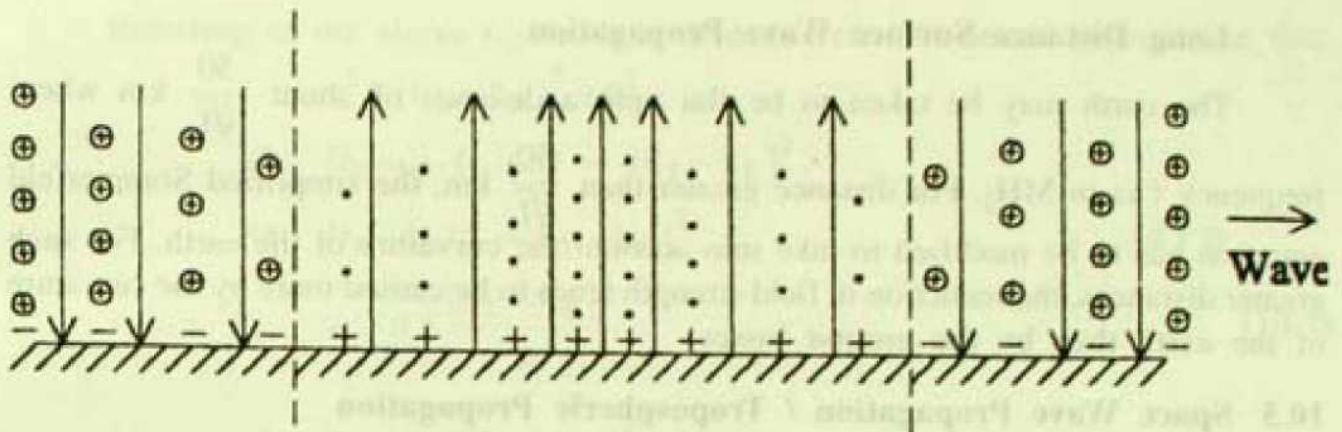
When both the transmitting antenna and the receiving antenna are at the ground surface then the direct and the ground reflected components of the space wave are equal in magnitude but opposite in phase. Thus the component cancel each other. Only the surface wave may be present. Such is the condition of ground wave propagation of broadcast at medium wave frequencies. As the height of the antennas, expressed in wave lengths, is increased the amplitude of the space wave increases rapidly and when the heights equal to a few wave lengths or more, the space wave becomes the main contributor of the ground wave.

The amplitudes of the space wave and the surface wave are influenced by the following factors — (a) resistivity and dielectric constant of the earth, (b) the frequency of the radio wave, (c) the heights of the transmitting and receiving antennas i.e.  $h_T$  and  $h_R$  respectively, (d) curvature of the earth, (e) the distance,  $d$ , between the transmitting and receiving antennas, and (f) the variation of the refractive index of the earth's atmosphere with height. The variation of the refractive index with height causes the path of propagation to be slightly curved instead of straight path, the curvature being in the same direction of the curvature of the earth.

#### 10.4 Surface Wave Propagation

The surface wave glides over the surface of the earth. This wave is vertically polarized, any horizontal component of electric field in contact with the surface of the earth will be short circuited. The surface wave induces, as shown in Fig. 10.3, charges in the earth which travel with the wave and thus constitute a current.





- : magnetic field coming out of paper
- + : magnetic field going into the paper

Fig. 10.3 Surface wave propagation

As the surface wave passes over the surface of the earth it is attenuated due to absorption of energy by the earth. Energy lost in this way is partly replenished by diffraction of additional energy downward from the portion of the wave present above the immediate surface of the earth.

When the transmitting and receiving antennas are close to the ground, the space wave is negligibly small and only the surface wave is important. Thus the daytime propagation of broadcast in medium wave takes place due to surface wave.

The surface wave propagation may again be divided into two categories:

- a. Short distance surface wave propagation assuming flat earth.
- b. Long distance surface wave propagation considering the curvature of the earth.

When the distance of the receiving antenna from the transmitting antenna is short enough, the curvature of the surface of the earth may be neglected; the earth may be assumed to be flat in the intervening distance. If the heights of the transmitting and receiving antenna expressed in wave length of the radio wave is low, then the surface wave alone may be considered. The field strength for surface wave propagation for a flat earth is given by the *Sommerfeld equation*

$$E = \frac{E_0}{d} A$$

Where  $E_0$  is the field strength at unit distance at the surface of the earth neglecting earth losses,

$d$  is the distance of the transmitting antenna,

and  $A$  is a factor which accounts for the losses caused by the earth.



## Long Distance Surface Wave Propagation

The earth may be taken to be flat upto a distance of about  $\frac{50}{\sqrt[3]{f}}$  km where frequency  $f$  is in  $\text{MHz}$ . For distance greater than  $\frac{50}{\sqrt[3]{f}}$  km, the simplified Sommerfeld equation has to be modified to take into account the curvature of the earth. For such greater distances, the reduction in field strength tends to be caused more by the curvature of the earth than by the ground losses.

### 10.5 Space Wave Propagation / Tropospheric Propagation

At radio frequencies above about 30  $\text{MHz}$ , the ionosphere is not able to refract energy to the earth, while the surface wave is attenuated to negligible amplitude within few hundred metres. Useful propagation can, however, be achieved at these radio frequencies by means of the space wave travelling between elevated transmitting and receiving antennas.

*Space wave propagation over ideal flat earth*

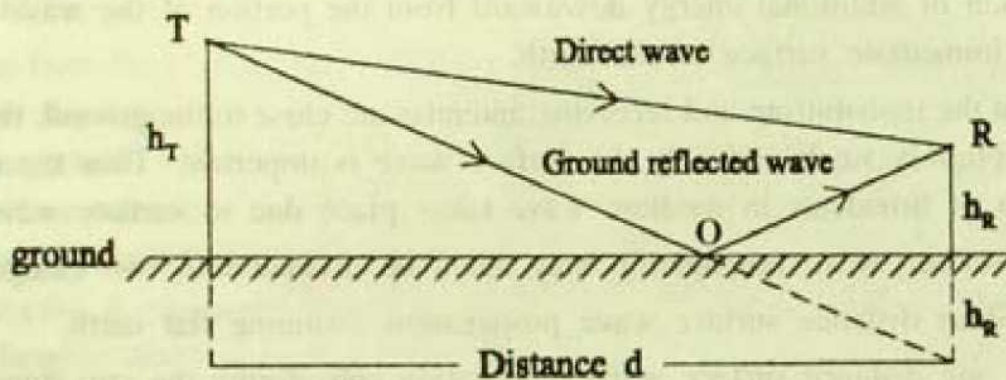


Fig. 10.4

Neglecting the curvature of the earth, the space wave propagates in the way shown in Figure 10.4 above. Here energy reaches the receiver in two ways (1) by a ray travelling directly between transmitting and receiving antennas over the path  $TR$  and (2) by a ray travelling over path  $TOR$  after reflection from the surface of the earth. The field strength at the receiving antenna  $R$  is the vector sum of the fields represented by these two rays. As the individual rays travel through space, they undergo negligible attenuation other than that caused by spreading. So, each wave considered alone has a field strength that is inversely proportional to the distance from the transmitter.

When the distance between transmitting and receiving antennas is much greater than the height of the antennas, the angle of incidence of the ray  $TO$  at the surface of the earth will be small. The reflection of radio wave at  $O$ , at this low angle of incidence, takes place without change of magnitude but with phase reversal. Under these circumstances the two radio waves at the receiving point  $R$  have equal amplitude, but will differ in phase, in general.



Referring to the above figure, it is evident from the dotted construction that

$$r_1^2 = (h_T - h_R)^2 + d^2$$

$$\text{or } (r_1 - d)(r_1 + d) = (h_T - h_R)^2$$

$$\text{or } (r_1 - d) 2d = (h_T - h_R)^2 \quad \therefore d = r_1$$

$$\therefore r_1 = d + \frac{(h_T - h_R)^2}{2d} \quad (10.3)$$

$$\text{Again, } r_2^2 = (h_T + h_R)^2 + d^2$$

$$\therefore r_2 = d + \frac{(h_T + h_R)^2}{2d} \quad (10.4)$$

The difference in path lengths between the ground reflected and direct ray is

$$\begin{aligned} r_2 - r_1 &= \frac{(h_T + h_R)^2}{2d} - \frac{(h_T - h_R)^2}{2d} \\ &= 2 \frac{h_T h_R}{d} \end{aligned} \quad (10.5)$$

The phase difference corresponding to this path difference is

$$\begin{aligned} &\frac{2\pi}{\lambda} \cdot \frac{2 h_T h_R}{d} \\ &= 4\pi \frac{h_T h_R}{\lambda d} \text{ radians} \end{aligned} \quad (10.6)$$

It is because of this angle for which the direct and reflected rays fail to cancel.

The resultant of the two waves having phase difference  $4\pi \frac{h_T h_R}{\lambda d}$  will be  $2 \sin \frac{2\pi h_T h_R}{\lambda d}$  times the amplitude of each wave.

Assuming  $d \gg h_T$ , and  $d \gg h_R$ , the field strength  $E$  at the receiver due to each wave can be expressed as

$$E = \frac{E_0}{d} \quad (10.7)$$

where  $E_0$  is the field intensity at unit distance produced by the transmitting antenna in the desired direction when the earth is absent (i.e. strength of the direct ray at unit distance)

and  $d$  is the distance between transmitting and receiving antenna.

The resultant field strength  $E_R$  of the space wave be due to direct and ground reflected rays will therefore.

$$E_R = 2 \frac{E_o}{d} \sin \frac{2\pi h_T h_R}{\lambda d} \quad (10.8)$$

when  $d$  is small, there is sinusoidal variation of the field strength, but when  $d$  is large such that  $\frac{2\pi h_T h_R}{\lambda d}$  is less than 0.5, the sine of the angle can be replaced by the angle. In this case the resultant field strength  $E_R$  becomes

$$E_R = 2 \frac{E_o}{d} \cdot \sin 2\pi \frac{h_T h_R}{\lambda d} = 2 \frac{E_o}{d} 2\pi \frac{h_T h_R}{\lambda d} = E_o \frac{4\pi h_T h_R}{\lambda d^2} \quad (10.9)$$

Typical curves showing the variation of field strength with distance as evident from the above expressions are shown in Figure 10.5.

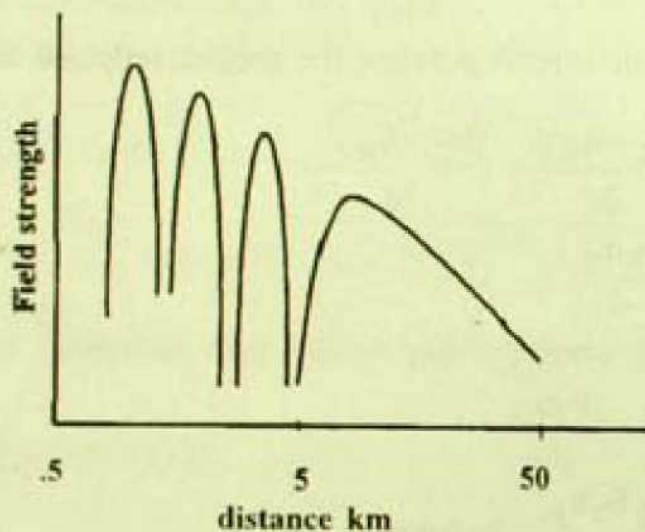


Fig. 10.5 Variation of field strength

It will be seen that for distance less than the value  $d^1$  that makes the angle  $2\pi \frac{h_T h_R}{\lambda d}$  greater than  $\frac{\pi}{6}$ , the field strength oscillates about the value  $\frac{E_o}{d}$  that corresponds to the strength of the direct ray TR. This is sometimes called the free space wave. For a perfectly conducting earth the maximum amplitude of these oscillations is twice the free space value and occurs at distances so related to the antenna heights that the direct and the ground reflected rays add in phase. The minima have zero amplitude in the case of perfectly reflecting earth and occurs at distances such that the two component waves cancel each other.

Increasing the quantity  $\frac{h_T h_R}{\lambda}$ , either by increasing the heights of the transmitting or receiving antenna or both or by increasing the frequency of the radio wave will have the same effect as seen for small value of  $d$ , giving oscillatory nature of the signal.



### *Effect of the curvature of the earth*

The discussion given above assumes the earth to be flat. However, when the distance between transmitting and receiving antennas is large, it becomes necessary to take into account the curvature of the earth.

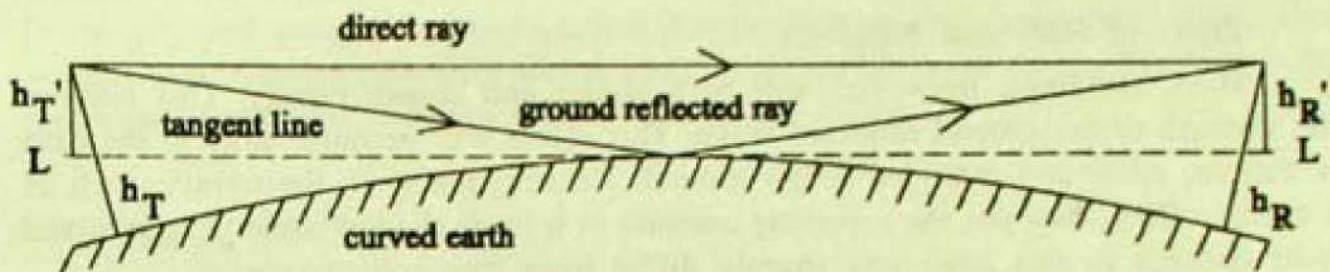


Fig. 10.6 Propagation over curved earth

When the receiving antenna is above the horizon of the curved earth, the geometry of propagation is shown in Figure. 10.6. The field strength at the receiving antenna will again be the vector sum of a direct ray and a ground reflected ray. But there is some change. Because of the curvature of the earth, the antenna heights to be taken into consideration will be the effective heights  $h'_T$  and  $h'_R$  respectively above the tangent line LL drawn at the point of reflection on the ground surface. Since these effective heights are less than actual heights, one effect of earth's curvature will be a change in the location and number of maxima and minima of the field strength curve.

Another consequence of earth's curvature arises from the fact that when a receiving antenna at moderate height is at a large distance from the transmitting antenna, it is below the radio horizon and so cannot be reached either by a direct or a ground reflected wave. This situation is shown in the Figure 10.7 where the shaded area is termed as the shadow zone or diffraction zone.

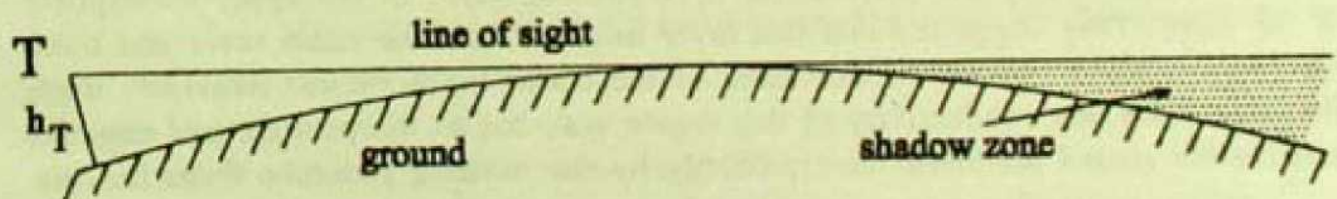


Fig. 10.7 Radio shadow zone

However, there are several mechanisms by which at least some energy from the transmitting antenna can reach a receiving antenna located in the shadow zone. Several factors appear to contribute to this result. The waves may be diffracted around the curve surface of the earth in the same way that sound wave bend around a corner.



However, the strength of the diffracted wave in the shadow zone depends on the roughness of the earth's surface. Over most land paths, the roughness of the ground is large enough to cause the field in the shadow zone to be considerably large. Moreover, turbulence in the troposphere gives rise to small irregularities in the refractive index which cause energy from rays to be scattered to the shadow zone.

#### *Effect of Hills and Buildings*

Hills, buildings, trees, etc. will both scatter and absorb energy. This reduces the strength of the ground reflected wave. This effect will be quite large in the case of forests, cities and extended rough ground. An irregularity in the terrain, such as a hill or valley, may put the receiving antenna in a local shadow zone. The received field strength in this case may sharply differ from free zone reception.

### **10.6 Sky Wave Propagation or Ionospheric Propagation**

Radio waves in the short wave range radiating at a suitable angle with the ground travel through space and encounter the ionized region in the upper atmosphere. Under suitable conditions the incident radio waves bend downward due to refraction by the ionized region and again reach the ground at a distant point. Such a wave is called the sky wave and such a propagation of radio waves is called a sky wave propagation or ionospheric propagation. Long distance radio wave communication is possible through the sky wave propagation.

#### *Early history*

Marconi in 1901 was successful in sending wireless signals from Cornwall to Newfoundland across the Atlantic. It widened considerable thoughts in the scientific world regarding the mode of wireless propagation round the curved surface of the Atlantic. Since wireless waves are nothing but electromagnetic waves, physicists and mathematicians were naturally prompted to suppose it to be a phenomenon of diffraction. But their rigorous efforts showed that the diffraction effect was inadequate to explain the observed bending of the radio wave round the curved surface of the earth.

Kennelly in America and Heaviside in England in the year 1902, postulated almost simultaneously the presence of a conducting layer in the upper atmosphere of the earth. They suggested that this layer might deflect the radio wave and force the radio wave to follow the curvature of the earth. Furthermore, Heaviside made the suggestion that conductivity of this region was due to the positive and negative ions in the region produced most probably by the ionizing radiation from the sun.

Eccles, in the 1912, first pointed out the actual process by which the charged particles affect the propagation of radio waves through the ionosphere. Larmor in 1924 strengthened the Eccles' theory by some essential additions to that theory. This Eccles-Larmor theory is still regarded as the basic theory of propagation of radio wave in the ionosphere. This important theory has been supplemented by the magneto-ionic theory of propagation of radio wave developed by Appleton, Hartree and others.



### *The Ionosphere and its Layers*

By ionosphere of the earth is meant the upper part of the atmosphere where ionization is appreciable. This region of the earth's atmosphere absorb large amount of radiant energy from the sun. This causes ionization producing free electrons as well as positive and negative ions. But the ionisation in the ionosphere is not uniform. It is stratified in the form of layers. This stratification occurs due to the differences in the physical properties and chemical composition of the atmosphere at different heights and also because of unequal abilities of different gases in absorbing solar radiation of different frequencies.

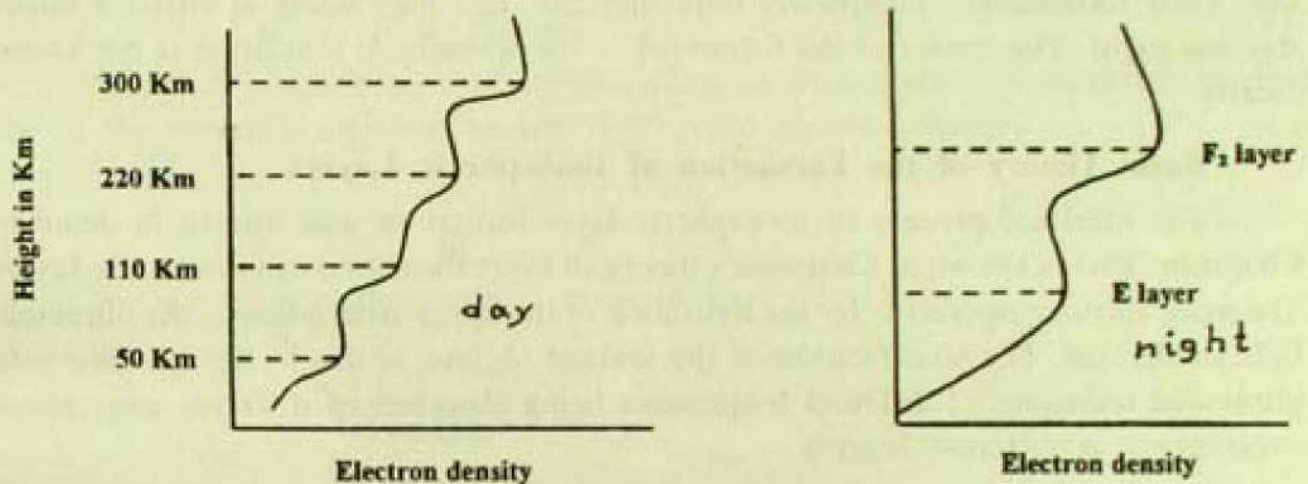


Fig. 10.8

The Figure 10.8 gives the variation of electron density during the day and night with height over the ground surface. The different levels of high electron density are called the layers of the ionosphere. There are three prominent maxima i.e. three prominent ionised layers in the daytime. These are E, F<sub>1</sub> and F<sub>2</sub>. There exists another ionised layer below the E layer known as D layer or D region. This D region offers attenuation to high frequency radio waves during daytime. This D layer exists between 50 to 90 km. This height varies from place to place and depends on the season of the year. The detailed composition of the D layer is not known with great certainty. The heights of highest electron density of the E layer and F<sub>1</sub> layer are more or less stable at values at about 110 km and 220 km respectively. These heights do not undergo appreciable diurnal and seasonal changes. On the other hand, the height of maximum electron density of the F<sub>2</sub> layer undergoes larger diurnal and seasonal variations. The typical height of the F<sub>2</sub> layer lie in the range 250 to 350 km. The diurnal and seasonal variations of the heights of maximum electron density of different layers are due to the variations in the composition and relevant temperature of the air at different heights and also due to different solar radiation. Thorough examinations reveal that ionization does not drop to zero between successive layers but actually drops to a value less than the adjacent maximum value on either side.



During nighttime the D layer vanishes due to the absence of solar radiation. The maximum electron density of the E layer depends on the solar ultraviolet radiation and so at night the maximum electron density of this E layer decreases uniformly with time. During night, the  $F_1$  and  $F_2$  layers of the ionosphere merge together to form one single layer which may be designated and usually designated as  $F_2$  layer.

In addition to these regular ionospheric layers, anomalous and sporadic ionisation in the form of very thin layer often appears in the E region at a height ranging from 90 to 130 km. This is known as 'Sporadic E'. This sporadic E layer often appears in the form of clouds of various sizes ranging from 1 km to several hundred km. Their formation is completely unpredictable and may occur at anytime during day and night. The causes of the formation of the sporadic E ionization is not known clearly.

### **Basic Theory of the Formation of Ionospheric Layers**

The idealised process of ionospheric layer formation was studied in detail by Chapman. This is known as Chapman's theory of layer formation of ionospheric layers. The main agency responsible for the ionization of the upper atmosphere is the ultraviolet light of the sun. The stratification of the ionized regions is due to the ionizing solar ultraviolet radiations of different frequencies being absorbed by different atmospheric constituents at different heights.

The process of formation of ionized stratum may be understood qualitatively very simply. Consider the earth to be enveloped by a single constituent gas. The density of the gas decreases with increasing height from the ground surface. If a beam of monochromatic radiation enters the earth's atmosphere from outside and is absorbed to produce ionization, then the rate of ion production will be maximum at a certain height. Because, the rate of absorption at any height is controlled by two factors — the intensity of the incident radiation and the density of the absorbing gas. As the ionizing radiation enters the atmosphere from above, the intensity of ionizing radiation decreases due to absorption with decreasing height from the ground surface while the density of the absorbing gas increases with decreasing height. The two opposing factors thus combine to produce maximum ionization at a height which is determined by the co-efficient of absorption of the gas for the particular radiation as well as by the density of the gas. Due to the different gases present in the atmosphere and due to the ionizing radiation of different frequencies, various ionospheric layers mentioned earlier, are formed.

### **Propagation of Radio Wave through Ionosphere**

As mentioned earlier, there are electrons, positive ions and negative ions in the ionosphere due to ionization. Out of them, the electrons are most affected by an electric field due to their low mass. In general whenever an electromagnetic wave passes through the ionosphere, the electric field of the radio wave exerts an appreciable



force on the electron. These electrons oscillate sinusoidally along paths parallel to the electric field of the wave constituting an alternating current proportional to the velocity of vibration. The resulting current is inductive, lagging behind the electric field by  $90^\circ$ .

#### Deduction

Let the electric field due to the incident radio wave which acts across the opposite face of a cubic metre of space in the ionosphere be

$$E = E_m \sin \omega t \quad \text{volts/metre} \quad (10.11)$$

If  $e$  be the charge of an electron in coulomb, then force acting on each electron is

$$f = -e E \quad \text{Newton} \quad (10.11)$$

Assuming no collision, this electron will have an instantaneous velocity  $v$  metres/sec in the direction opposite to the field given by the relation,

$$-e E = m \frac{dv}{dt} \quad (10.12)$$

where  $m$  is the electron mass in kg

Integrating

$$\begin{aligned} v &= \frac{1}{m} \int -e E dt \\ &= -\frac{e}{m} \int E_m \sin \omega t dt \\ &= -\frac{e}{m} \frac{E_m}{\omega} (-\cos \omega t) \\ &= \frac{e}{m\omega} E_m \cos \omega t \end{aligned} \quad (10.13)$$

Let  $N$  be the electron density expressed in electrons/cubic metre at the place of interest. Then the instantaneous electric current constituted by these  $N$  electrons moving with instantaneous velocity  $v$  is given by

$$\begin{aligned} i_c &= -Nev \\ &= -\left(\frac{Ne^2}{m\omega}\right) E_m \cos \omega t \end{aligned} \quad (10.14)$$

The above current  $i_c$  lags behind the electric field  $E (= E_m \sin \omega t)$  by  $90^\circ$ . In addition to this inductive current, there is the capacitive current  $i_c$  which obviously leads the electric field by  $90^\circ$ .

The capacitance due to unit volume i.e. one cubic metre is

$$\epsilon_0 = 8.854 \times 10^{-12} \text{ farad} \quad (10.15)$$

The charge  $Q$  on the opposite faces of the unit cube is

$$Q = \epsilon_0 E \quad (10.16)$$

Hence the capacitive current  $i_c$  through this capacitance is given by,

$$\begin{aligned} i_c &= \frac{d}{dt}(\epsilon_0 E) \\ &= \epsilon_0 \frac{d}{dt} E_m \sin \omega t \\ &= \epsilon_0 \omega E_m \cos \omega t \end{aligned} \quad (10.17)$$

The total current  $i$  that flows through a cubic metre of space is given by,

$$\begin{aligned} i &= i_c + i_c \\ &= \omega \left\{ \epsilon_0 - \frac{Ne^2}{m\omega^2} \right\} E_m \cos \omega t \end{aligned}$$

Thus the inductive current  $i_c$ , subtract from the capacitive current  $i_c$ . The presence of free electrons in space results in the decrease in the space current effectively reducing the effective dielectric constant of the space to a value below  $\epsilon_0$  which was there in the absence of free electrons.

Accordingly, the effective value of the dielectric is given by,

$$\epsilon = \epsilon_0 - \frac{Ne^2}{m\omega^2} \quad (10.19)$$

This reduction in the effective dielectric constant produced by the presence of free electrons in the ionosphere bends the path of travel of the electromagnetic wave. The process of the bending of the path of the radio wave in the ionosphere may be understood with reference to the refractive index of the ionized region. The refractive index of the ionized region is given by,

$$\mu = \sqrt{\epsilon_r} \quad (10.20)$$

where  $\epsilon_r$  is the relative dielectric constant of the ionized region relative to that of the free space,  $\epsilon_0$ .

$$\begin{aligned} \therefore \mu &= \sqrt{\frac{\epsilon}{\epsilon_0}} \\ &= \sqrt{\frac{\epsilon_0 - \frac{Ne^2}{m\omega^2}}{\epsilon_0}} \end{aligned}$$



$$= \sqrt{1 - \frac{Nc^2}{\epsilon_0 m\omega^2}} \quad (10.21)$$

Substituting the values of charge  $e$ , and mass  $m$  of the electron and the value of  $\epsilon_0$

$$\mu = \sqrt{1 - \frac{81N}{f^2}} \quad (10.22)$$

where  $f$  is the frequency in KHz.

Thus the effective values of refractive index  $\mu$  will always be less than unity. The deviation of the refractive index from unity increases with the increase of the electron density  $N$  of the ionosphere and with the decrease of the frequency  $f$  of the passing radio wave. The ionized medium has a  $\mu$  always less than unity. Moreover, the phase velocity of the radio wave in the ionosphere is always greater than the velocity of light because —

$$\text{phase velocity} = \frac{\text{velocity of light}}{\mu}$$

This phase velocity is higher at higher electron density  $N$  due to the lowering of  $\mu$  with higher values of electron density.

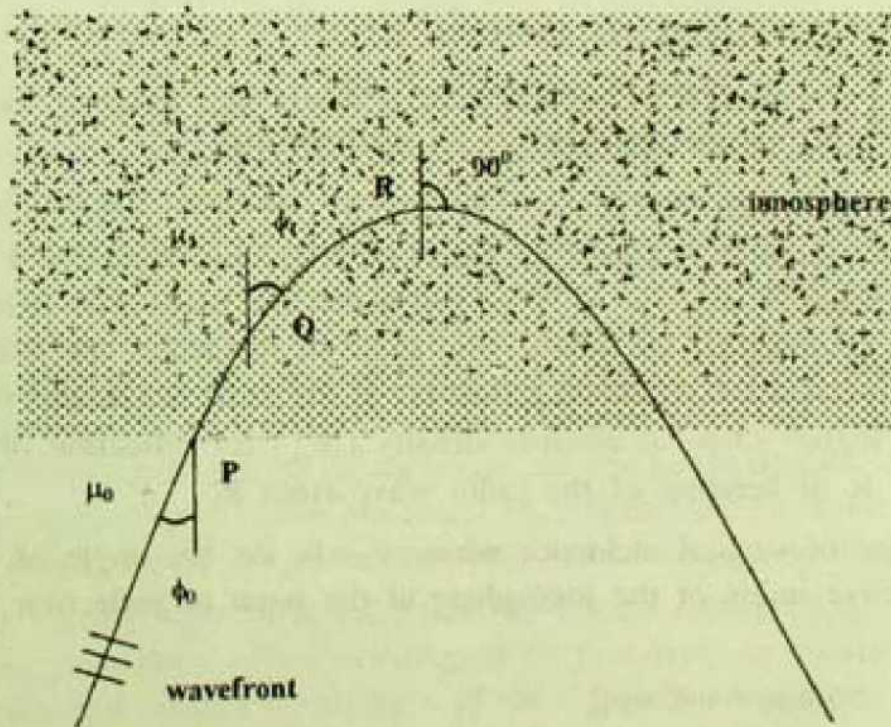


Fig. 10.9 Refraction of radio wave in the ionosphere

The phase velocity is a function of electron density and therefore when a radio wave enters the ionosphere, the edge of the wavefront lying in the region of higher electron density advances faster than the remaining part of the wavefront lying in the region of lower electron density. The path of travel of the radio wave is normal to the wavefront and therefore it bends. The bending of a radio wave produced by ionization follows the laws of optics. Thus the angle of the refracted ray at any point Q (Fig. 10.9) within the ionospheric layer and the refractive index at that point are related to those at other regions by Snell's Law which is

$$\mu_1 \sin \phi_1 = \mu_0 \sin \phi_0 \quad (10.23)$$

$$\begin{aligned} \text{at P } \left\{ \begin{array}{l} \mu_0 = \text{the refractive index of the free space} = 1 \\ \phi_0 = \text{the angle of incidence at the boundary of the ionosphere} \end{array} \right. \\ \text{at Q } \left\{ \begin{array}{l} \mu_1 = \text{the refractive index} \\ \phi_1 = \text{the angle of refraction} \end{array} \right. \end{aligned}$$

Assuming that the radio wave is entering the ionized region from downward, it encounters increasing electron density. So, the path of propagation of the radio wave bends more and more, until at some point R where the electron density is so large that the angle of refraction becomes  $90^\circ$ . From the point R, the radio wave is reflected and travels downwards. R is the highest point reached by the radio wave when incident at an angle  $\phi_0$ .

$$\therefore \mu_0 \sin \phi_0 = \mu_m \sin \phi_m \quad (10.24)$$

$$\begin{aligned} \text{at R } \left\{ \begin{array}{l} \mu_m = \text{refractive index} \\ \phi_m = \text{angle of refraction} = 90^\circ \\ N_m = \text{electron density in electrons/m}^3. \end{array} \right. \\ \therefore \sin \phi_0 = \mu_m \quad (10.25) \end{aligned}$$

Usually the point R is called the point of reflection. Actually it is the point at which the refraction angle is so large ( $90^\circ$ ) that the wave gets bent earthward. Moreover, it is evident that the smaller the angle of incidence, the smaller will be the value of  $\mu$  needed for so called reflection of the radio wave and smaller value of  $\mu$  demands higher value of electron density ( $N_m$ ) for reflection of radio wave from the point R or bending of the radio wave from R.

In the case of vertical incidence when  $\phi_0 = 0$ , i.e. the angle of incidence is zero, the refractive index of the ionosphere at the point of reflection must reduce to zero.

$$\begin{aligned} \sin \phi_0 = \sin 0 = \mu_m \\ \text{But } \mu_m = 0 = \sqrt{1 - \frac{81N_m}{f^2}} \quad (10.26) \end{aligned}$$



where  $N_m$  is the corresponding electrons density in electrons/m<sup>3</sup> and  $f$  is the frequency of the radio wave in kHz.

Therefore, the radio wave with vertical incidence penetrates the ionosphere until it reaches a point R, where the refractive index reduces to zero due to the appropriate electron density  $N_m$  there corresponding to the frequency of the incident radio wave.

Now, suppose that the maximum electron density of an ionospheric layer is  $N$  electrons/cubic metre. For vertical incidence where  $\phi_0 = 0$ , the highest frequency which can be reflected by this layer is one for which the refractive index of the layer at the point of maximum electron density just equals zero. This highest frequency  $f_c$  is therefore given by,

$$0 = \sqrt{1 - \frac{81N}{f_c^2}} \quad (10.27)$$

$$\text{or } f_c = \sqrt{81N} = 9\sqrt{N} \quad (10.28)$$

where  $f_c$  is in kHz

and  $N$  in electrons/metre cube.

The frequency  $f_c$  is called the critical frequency of the layer. This critical frequency of a layer may be defined as the highest frequency of a radio wave reflected from that layer when the wave is incident normally on that ionospheric layer. Radio waves having frequencies equal to or less than the critical frequency of a layer will definitely be reflected from that layer irrespective of the angle of incidence.

### Virtual Heights and Critical Frequencies of Layers

The virtual height of an ionospheric layer is the height to which a radio wave sent vertically would reach and come back to the ground travelling with the velocity of light taking the same total time as actually taken by the radio wave travelling with different velocities along its path of travel. The virtual height is higher than the true height of reflection because the velocity of propagation of the radio wave in the ionized region gets reduced below the velocity of light due to exchange of energy between the wave and the electrons. The amount of reduction in velocity and hence the difference between the virtual height and true height depends on the distribution of electrons in the layer below the level of reflection. The difference in the two heights is usually small but sometimes may become large upto 100 km or so.

It has already been mentioned that the critical frequency of an ionized layer is the highest frequency which is reflected by that layer at vertical incidence. The virtual heights and critical frequencies of the different ionospheric layers undergo diurnal and seasonal variations. The variation of the critical frequency is more prominent and very important for radio wave propagation. Their variations are shown in Figure 10.10a and b.



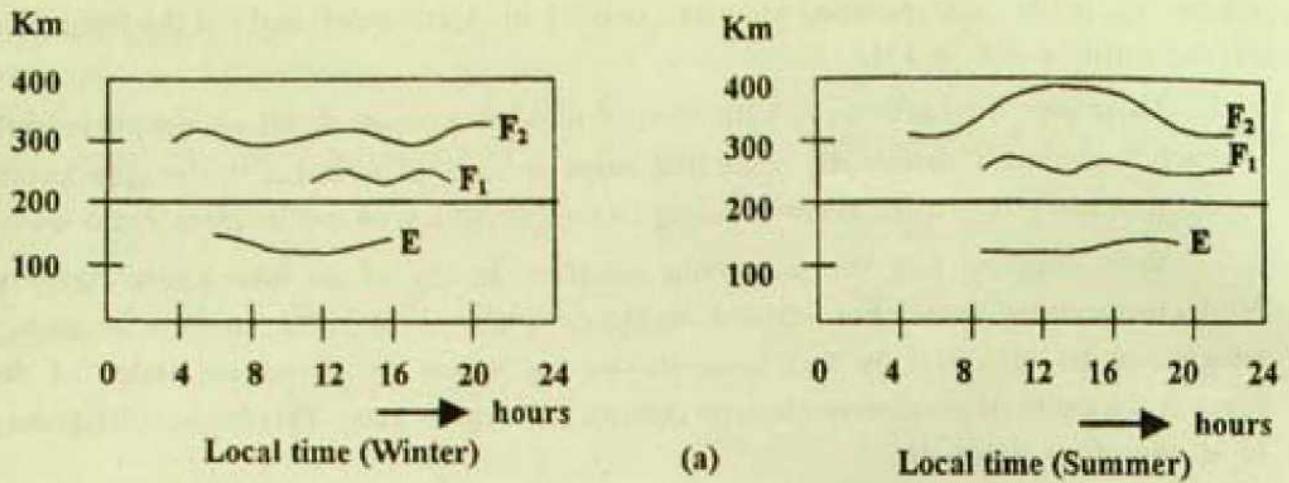


Fig. 10.10a. Diurnal variation of virtual height

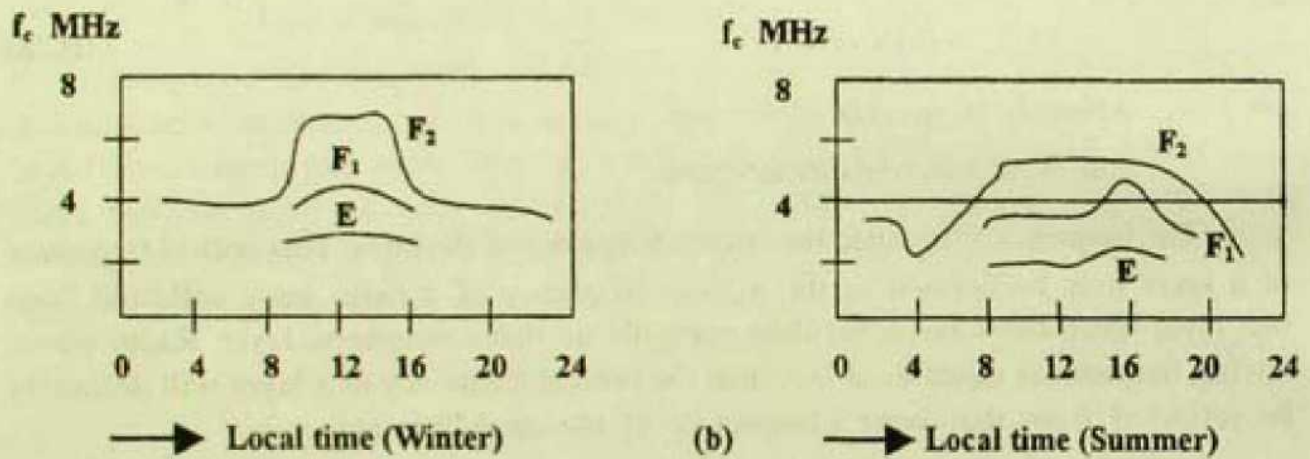


Fig. 10.10b. Diurnal variation of critical frequency

The figures show the nature of the diurnal variation of the virtual heights and critical frequencies of ionospheric layers averaged over winter and summer months for a location in temperate climate.

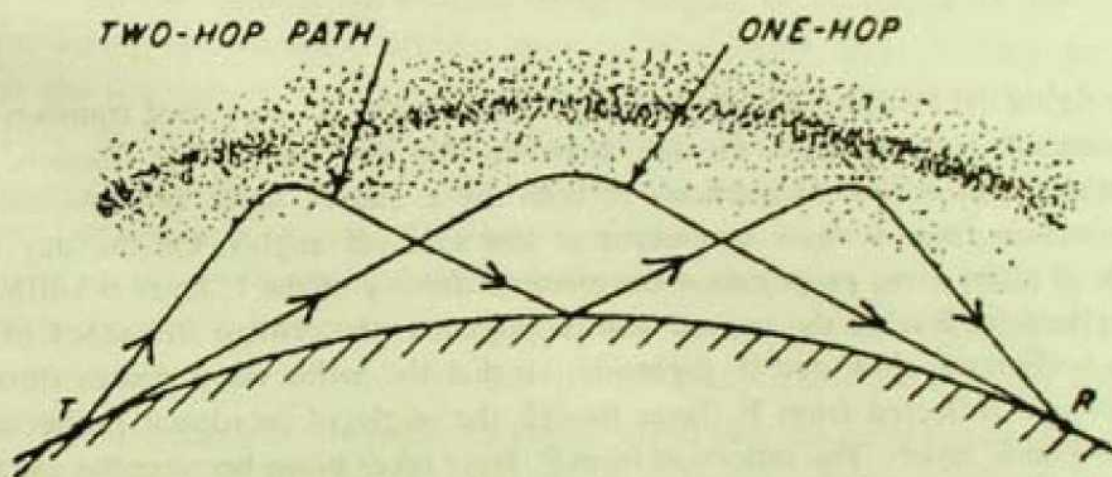
The critical frequencies of the E and F<sub>1</sub> layer follow a regular diurnal cycle. The critical frequency of the F<sub>2</sub> layer undergoes large diurnal as well as seasonal variations. Moreover, the critical frequencies of the regular layers— E, F<sub>1</sub> and F<sub>2</sub> — decrease considerably during nighttime due to decreasing ionization in the absence of solar radiation.

### Oblique Incidence and Multihop Propagation

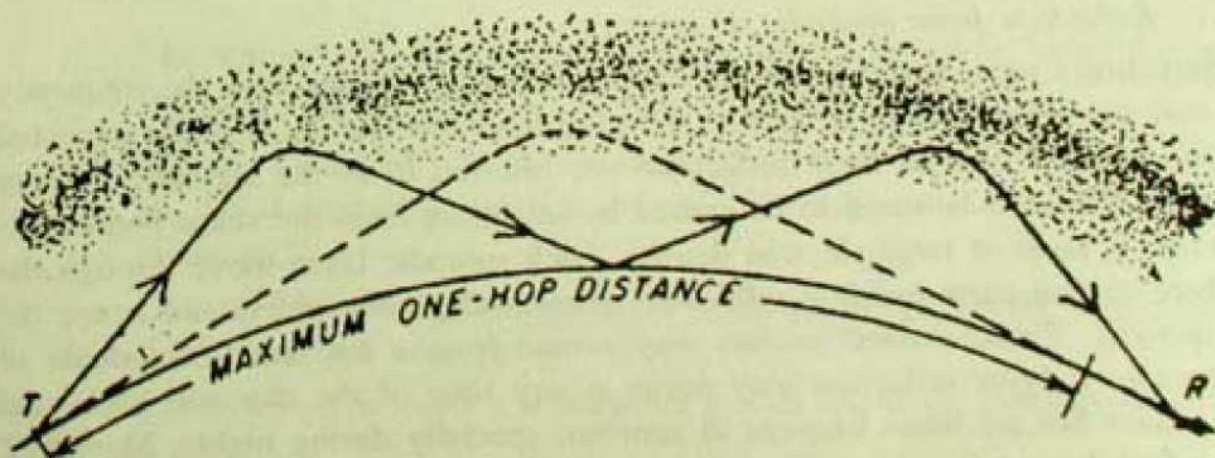
Vertical incidence of radio wave is of no practical importance from propagation point view from the transmitter to the receiver which are at a considerable distance from each other. Oblique incidence of the radio wave on the ionosphere is invariably used for broadcasting. The maximum frequency that can be reflected by a layer on



vertical incidence was designated as the critical frequency. But on oblique incidence radio waves having frequencies appreciably larger than critical frequency may be reflected by the same layer. Their relation is given by a law known as secant law. Sky waves using short waves use frequencies much larger than the critical frequency of the reflecting layers. Moreover, the longest single-hop possible for radio wave considering the curvature of the earth corresponds to a radio wave that leaves the transmitter tangent to the earth's surface. The earth curvature is such that for typical virtual heights of the ionospheric layers this maximum one hop distances are about 2000 km for E-layer and about 4000 km for  $F_2$  layer. At larger distances, radio wave transmission is possible only by means of two or more hops depending on the total distance. Propagation will then be affected by condition in the ionosphere at each reflection point along the path. These conditions will in general be different over east-west paths because of time differences.

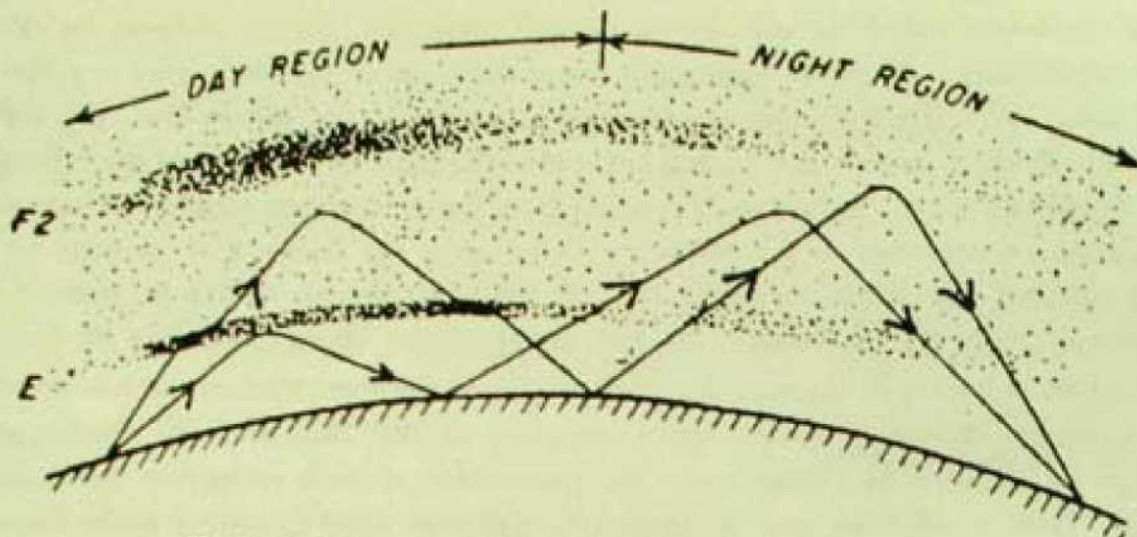


A. Single hop and double hop transmission



B. Typical example of two hop transmission





C. Multihop transmission involving E and  $F_2$  layer

Fig. 10.11

The figure 10.11 illustrates the propagation in hops. An east-west transmission path is assumed which crosses the line between day and night. The frequency is chosen greater than critical frequencies of both the E and  $F_2$  layer, but low-enough so that reflection from E layer can occur at low take-off angles. On the day side of the path of radio wave propagation the electron density of the E layer is sufficient to cause reflection. But on the second hop propagation, the critical frequency of the E layer is sufficiently low due to nightside, so that the radio wave passes through the E layer and reflected from  $F_2$  layer though the angle of incidence is the same on the ionospheric layers. The reflection from  $F_2$  layer takes place because the electron density of the  $F_2$  layer at night is higher than the electron density of the E layer at day.

### Abnormal behaviour of the Ionosphere

#### i) *Reflection from Sporadic E layer :*

Sometimes radio waves at frequencies considerably higher than the frequency which may be reflected from E layer is reflected from that region. It is not expected from the normal E region. Such reflections are referred to as the sporadic E layer reflections. These are believed to be caused by reflection from the sharp boundaries of sporadic E layer of large electron density. Such sporadic layer move through the ionosphere so that corresponding reflections come and go irregularly and hence the name sporadic. These ionized patches may extend from a few km to hundreds of km. Sporadic E layer reflection may occur at any time of the day and night and in any season but are more frequent in summer, specially during nights. Moreover, the sporadic E layer reflection appear to be more frequent during the period of maximum sun-spot activity.



## ii) *Radio Fade out : Dellinger Effect*

Sometimes ionospheric radio wave suddenly fadeout or disappear. This phenomenon is called the Dellinger Effect and is caused by bursts of ionizing radiation from the sun. Such high and sudden radiation causes abnormally high ionization in the ionosphere below the E layer. This causes heavy absorption of the radio wave passing through the region. The radio fade out sets in within a minute and may last for any period from a few minutes to several hours. Obviously the phenomenon does not occur at night. It occurs during daytime simultaneously throughout that part of the globe facing the sun.

## iii) *Ionospheric Storms :*

Ionospheric storm is the phenomon associated with the poor radio transmission at frequencies above 500 kHz lasting for one or more days. Usually it is accompanied by geo-magnetic storm. An ionospheric storm begin with violent turbulence of the entire ionosphere causing destruction of the normal stratification of the ionosphere and producing small ionized regions moving in irregular ways. During the phase of the storm, ionospheric layers with specific virtual heights and critical frequencies disappear. In the initial phase, the storm affects the auroral zone violently. In the second phase of the storm, the effects initiated in the auroral region gradually spread to lower latitudes depending on the strength of the storm. In the last phase, the ionosphere gradually returns to the normal taking long time, sometimes as large as several days. The magnetic storms during day and night without distinction and frequency of occurrence of storm closely follows the solar activity and thus follow the 11 year solar cycle.

The ionospheric storm produces greater effect near the polar regions but negligible effect near the equator. The turbulent phase generally results in regions within  $20^\circ$  surrounding geomagnetic pole. The second phase is typically found in the temperate zone and occurs several hours after the initial phase.

## 10.7 Exercises

1. Electromagnetic waves are said to be transverse. What does this mean?
2. Explain what is meant by isotropic radiator and isotropic medium.
3. What is meant by linear polarisation of an e.m. wave? What do you mean by the statement — an e.m. wave is horizontally polarised or vertically polarised?
4. Give a table showing various radio frequency range that are in use. Discuss their means of propagation.
5. Comment on the earth's atmosphere and its role in the propagation of radio waves.
6. Describe ground wave propagation. Explain the expression of its field strength at a distance from the transmitter.

7. What is the ionosphere? What are the different ionospheric layers? How they are formed?
8. Describe the different ionospheric layers with reference to their heights. Describe their role on radio wave propagation.
9. Why short wave propagation is generally better at night than during the day?
10. What do you mean by sky wave? Explain it.
11. Describe briefly the mechanism by which electromagnetic waves are bent back by a layer of the ionosphere.
12. Show, with suitable diagram, what happens when the angle of incidence of a radio wave is brought closer and closer to vertical in the case of sky wave propagation.
13. Give definitions and briefly describe the following terms in connection with sky wave propagation. — (a) critical frequency, (b) virtual height.
14. Derive expression and explain clearly the process of reflection of a radio wave by the ionized media of the ionospheric layers.
15. What is a sporadic E layer? What is its effect on the propagation of radio wave?
16. Explain clearly the process of space wave propagation.
17. What is the radio horizon? How does it differ from the optical horizon? Discuss it.
18. What is the effect of the curvature of the earth in radio wave propagation?
19. What is single hop and multihop propagation? Explain clearly east to west and west to east propagation.
20. Mention the three important methods of radio wave propagation. Summarize briefly the mechanism of their propagation.

## 10.8 References

1. Electronic and Radio Engineering — F. E. Terman, McGraw Hill Book Company, INC.
2. Electronic Communication — D. Roddy & J. Coolen, Prentice Hall of India Pvt. Ltd.
3. Electronic Communication Systems — G. Kennedy, McGraw Hill Kogakusha Ltd.
4. Applied Electronics — G. K. Mithal, Khanna Publishers, Delhi.



## Chapter 11

### Transmission of Electromagnetic Wave

#### 11.1 Modulation and Demodulation

The necessity of modulation arose out of human desire for communication of intelligence or signal over long distances. Range of communication through sound wave is extremely limited.

In the MW (medium wave) radio broadcasting the maximum audio frequency allowed is 5 kHz. Thus from simple consideration one channel or one programme will occupy frequency upto 5 kHz. Only one such message may normally be sent over a pair of wires. If two such programmes are sent simultaneously on the same pair of wires, the two messages will get mixed up and the information is lost. Such an undesirable situation may, however, be avoided by shifting the second message in frequency spectrum away from the first message. Thus the second message initially occupying frequency spectrum upto 5 kHz may be shifted in the range 5 kHz to 10 kHz. The two signals then do not get mixed. Several such messages occupying different specific ranges without overlapping may then be sent over the same pair of wires. This is in fact the principle of carrier current telephony. If however the message is shifted to a relatively higher frequency spectrum, say at 1 MHz to 1.005 MHz then it is possible to radiate this signal through space without requiring any telephone line.

The term modulation means regulation or adjustment and specifically in the case of telecommunications it means to regulate some parameters of a high frequency by means of a lower frequency signal. The need for modulation first arose in the radio transmission of low frequency audio signal. It is a fact that for efficient radiation antenna dimension had to be of the same order as the wavelength of the signal being transmitted. The frequency  $f$  and the wavelength  $\lambda$  of an electromagnetic wave are related to the phase velocity  $v$  by

$$f\lambda = v$$

Take the case of an audio frequency signal at 1 kHz. Since electromagnetic waves travel in free space with velocity of light, hence

$$\lambda = \frac{3 \times 10^8}{1000} \text{ m} = 300 \text{ km}$$

$$= 188 \text{ miles}$$

Obviously it is impractical to build antennas of this size.

The problem is to overcome this problem by using the low audio frequency signals to modulate a higher frequency signal say at 1 MHz when the wavelength reduces to

$$\lambda = \frac{3 \times 10^8}{10^6} \text{ m} = 300 \text{ m}$$

Radio signals at this wavelength may be radiated efficiently with suitable antenna.

This is in brief the working of radio communication. The working of radio communication as well as the carrier current telephony requires shifting or translating the signals from the original position in the frequency spectrum to a higher position in the frequency spectrum.

## Modulation

Modulation is defined as the process by which some characteristics, usually, amplitude, frequency or phase, of a voltage is varied in accordance with the instantaneous value of some other voltage called the modulating voltage. The term 'carrier' is applied to the voltage whose characteristics is varied and the term "modulation voltage" or 'modulating voltage' is used for the voltage by which the variation takes place. Usually the modulation frequency is considerably lower than the carrier frequency.

Let the carrier voltage be represented by

$$e_c = E_c \cos (\omega_c t + \theta)$$

Where  $\theta$  = initial phase or epoch

$t$  = time

$\omega_c$  = angular frequency in radians/sec

$$= 2\pi f_c$$

Where  $f_c$  = the frequency of the carrier voltage in Hz.

The modulation process then may consist in varying one of the following three quantities of the carrier voltage

(a) amplitude  $E_c$

(b) frequency  $\omega_c$

(c) phase angle  $\theta$

The variation of the carrier voltage will be in accordance with some function of the instantaneous value of the modulating voltage.

Accordingly modulation process may be classified as

(a) amplitude modulation, AM

(b) frequency modulation, FM

(c) phase modulation, PM.

It depends on whether the amplitude  $E_c$ , frequency  $\omega$  or the phase angle  $\theta$  of the



carrier voltage is varied according to some function of the instantaneous value of the modulating voltage. Sometimes the frequency modulation and phase modulation are placed in one group and combinedly called the 'angle modulations'.

## 11.2 Amplitude Modulation

In amplitude modulation, the amplitude of the carrier voltage varies in accordance with the instantaneous value of the modulating voltage.

Let the modulating voltage be represented by

$$e_m = E_m \cos \omega_m t$$

where  $E_m$  = amplitude of the signal

$$\omega_m = \text{angular frequency in radians/sec}$$

Let the carrier voltage be represented by

$$e_c = E_c \cos \omega_c t \text{ [assuming } \theta = 0]$$

For convenience of calculation, initial phase  $\theta$  or epoch is taken as zero, since it does not play any part in the modulation process. This however, does not reduce the generality of the expression.

In the process of amplitude modulation, the amplitude of the carrier voltage does not remain constant but varies according to the instantaneous value of the modulating voltage. This variation of amplitude is proportional to the modulating voltage i.e.

$$\text{Variation in carrier amplitude} \propto E_m \cos \omega_m t = K_a E_m \cos \omega_m t$$

where  $K_a$  is the proportionality constant.

Therefore the instantaneous amplitude of the modulated voltage is given by

$$E_t = E_c + K_a E_m \cos \omega_m t$$

And the instantaneous value of the modulated carrier voltage is

$$e = E_t \cos \omega_c t$$

$$= (E_c + K_a E_m \cos \omega_m t) \cos \omega_c t$$

$$= E_c \left( 1 + \frac{K_a E_m}{E_c} \cos \omega_m t \right) \cos \omega_c t$$

$$= E_c (1 + m_a \cos \omega_m t) \cos \omega_c t$$

$$\text{where } m_a = \frac{K_a E_m}{E_c}$$

$m_a$  is called the 'modulation index' or "modulation factor" or 'depth of modulation'.

The percentage modulation is given by,

$$100 \times m_a = \% \text{ modulation.}$$

The waveforms of the carrier voltage  $E_c$ , modulating voltage  $e_m$  and modulated voltage  $e$  are given in A, B and C respectively. The crest of the modulated voltage corresponds to the crest of the modulating voltage whereas the trough of the modulating voltage coincides with the trough of the modulated waveform (Fig. 11.1).

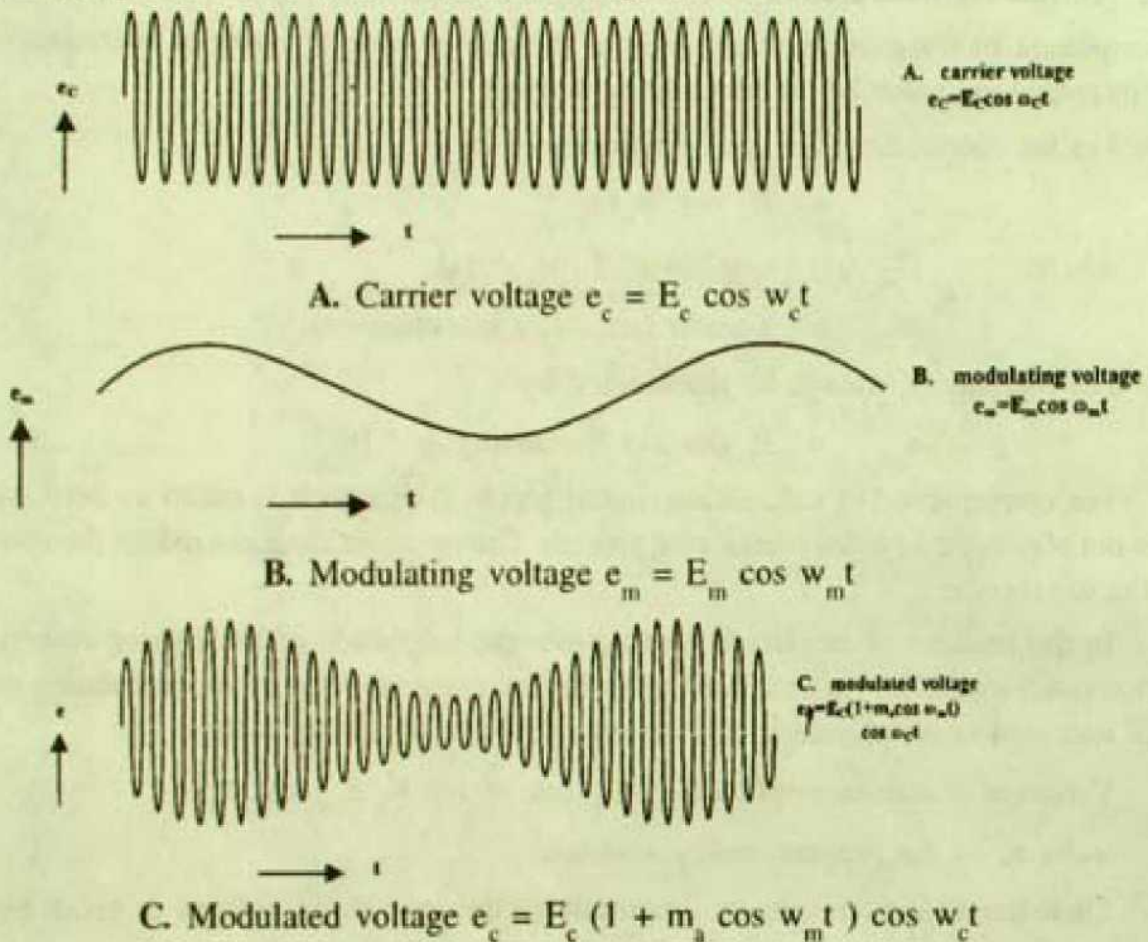


Fig. 11.1 Waveform of AM wave

The instantaneous amplitude of the modulated voltage is

$$E_t = E_c (1 + m_a \cos \omega_m t)$$

The maximum value of  $E_t = E_c (1 + m_a)$  when  $\cos \omega_m t = 1$   
 $= E_{\max}$  (say)

Similarly the minimum value of  $E_t = E_c (1 - m_a) = \text{say, } E_{\min}$

From these two expression, the value of the modulation index,  $m_a$  comes out to be

$$m_a = \frac{E_{\max} - E_c}{E_c}$$

$$\text{and also } m_a = \frac{E_c - E_{\min}}{E_c}$$

Further adding and subtracting these expressions



$$E_{\max} + E_{\min} = 2E_c$$

$$\text{and } E_{\max} - E_{\min} = 2E_c m_a$$

$$\text{which gives } m_a = \frac{E_{\max} - E_{\min}}{E_{\max} + E_{\min}}$$

The waveform of the amplitude modulated wave can be observed and the value of modulation index,  $m_a$  can be calculated with the help of cathode ray oscilloscope.

### Frequency spectrum of AM wave

The amplitude modulated carrier voltage is given by

$$\begin{aligned} e &= E_c (1 + m_a \cos \omega_m t) \cos \omega_c t \\ &= E_c \cos \omega_c t + \frac{E_c m_a}{2} \{ \cos (\omega_c + \omega_m) t + \cos (\omega_c - \omega_m) t \} \\ &= E_c \cos \omega_c t + \frac{E_c m_a}{2} \cos (\omega_c + \omega_m) t + \frac{E_c m_a}{2} \cos (\omega_c - \omega_m) t \end{aligned}$$

It reveals that the sinusoidal carrier voltage  $e_c$  when amplitude modulated by a sinusoidal modulating voltage  $e_m$  consists of three frequency components,  $\omega_c$ ,  $\omega_c + \omega_m$  and  $\omega_c - \omega_m$ . Thus the amplitude modulated wave contains in addition to the carrier frequency, two other frequencies equal to the sum and difference of the carrier frequency and modulation frequency. The sum of the two frequencies ( $\omega_c + \omega_m$ ) is called the upper side frequency and the difference of the two frequencies ( $\omega_c - \omega_m$ ) is called the lower side frequency.

Their amplitudes are  $E_c$ ,  $\frac{E_c m_a}{2}$  and  $\frac{E_c m_a}{2}$  respectively. The lower side frequency term and the upper side frequency term are located in the frequency spectrum on either side of the carrier frequency  $\omega_c$ , at a frequency interval of  $\omega_m$ . Such a diagram is called a frequency spectrum (Fig. 11.2).

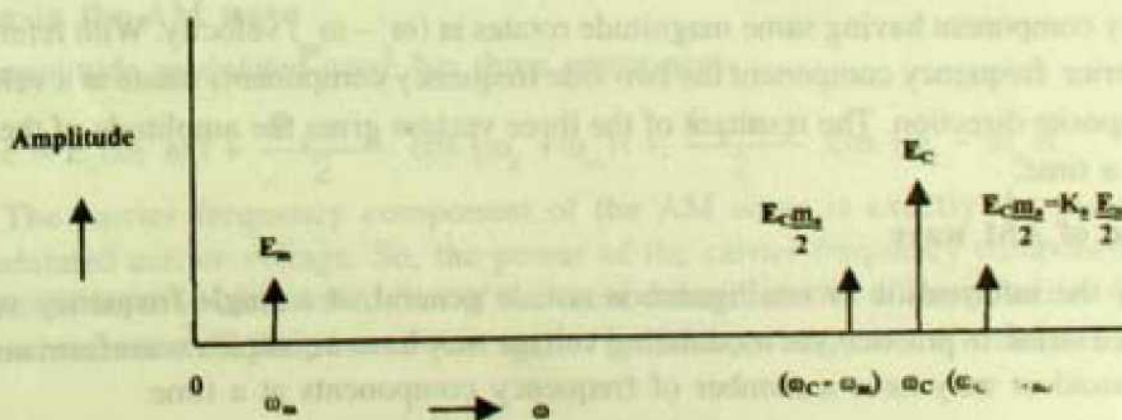


Fig. 11.2 Frequency spectrum of AM wave

The process of amplitude modulation shifts the information or intelligence from the low (audio) frequency position  $\omega_m$  to the high frequency position at  $(\omega_c + \omega_m)$  and  $(\omega_c - \omega_m)$  placed symmetrically around  $\omega_c$ . Each of these side frequency component voltage contains the complete information or intelligence originally contained in the signal before modulation at a low audio frequency. The intelligence occurs twice, one at  $\omega_c + \omega_m$  and the other at  $\omega_c - \omega_m$  though the original information or intelligence was at single frequency  $\omega_m$  only. The centrally placed voltage  $E_c \cos \omega_c t$  carried no information or intelligence but it helps in the transmission of intelligence.

The three components of a amplitude modulated wave viz.  $E_c \cos \omega_c t$ ,  $\frac{E_c m_a}{2} \cos (\omega_c + \omega_m)t$  and  $\frac{E_c m_a}{2} \cos (\omega_c - \omega_m)t$  at frequency  $\omega_c$ ,  $(\omega_c + \omega_m)$  and  $(\omega_c - \omega_m)$  respectively may be represented suitably in a vector diagrams (Fig. 11.3).

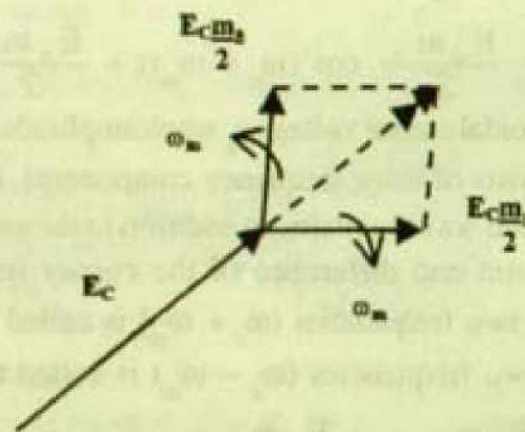


Fig. 11.3 Vector diagram

The carrier amplitude  $E_c$  rotates with angular velocity  $\omega_c$ . The upper side frequency component  $\frac{E_c m_a}{2} = \frac{K_a E_m}{2}$  rotates at an angular velocity  $(\omega_c + \omega_m)$  and the lower side frequency component having same magnitude rotates at  $(\omega_c - \omega_m)$  velocity. With reference to the carrier frequency component the two side frequency components rotate at a velocity  $\omega_m$  in opposite direction. The resultant of the three vectors gives the amplitude of the AM wave at a time.

### Sideband of AM wave

In reality the information or intelligence is not, in general, at a single frequency  $\omega_m$  as considered so far. In practice, the modulating voltage may have a complex waveform instead of a sinusoid or may have a number of frequency components at a time.

Each frequency component produces two side frequency components one placed above the carrier frequency and the other placed symmetrically below on both sides of the carrier frequency. Due to the closely separated frequencies in the modulating voltage, the modulated



voltage contains two groups of closely separated frequency on either side of the carrier frequency. The group of frequencies above the carrier frequency is called the upper sideband while the group of frequencies below the carrier frequency is called the lower sideband.

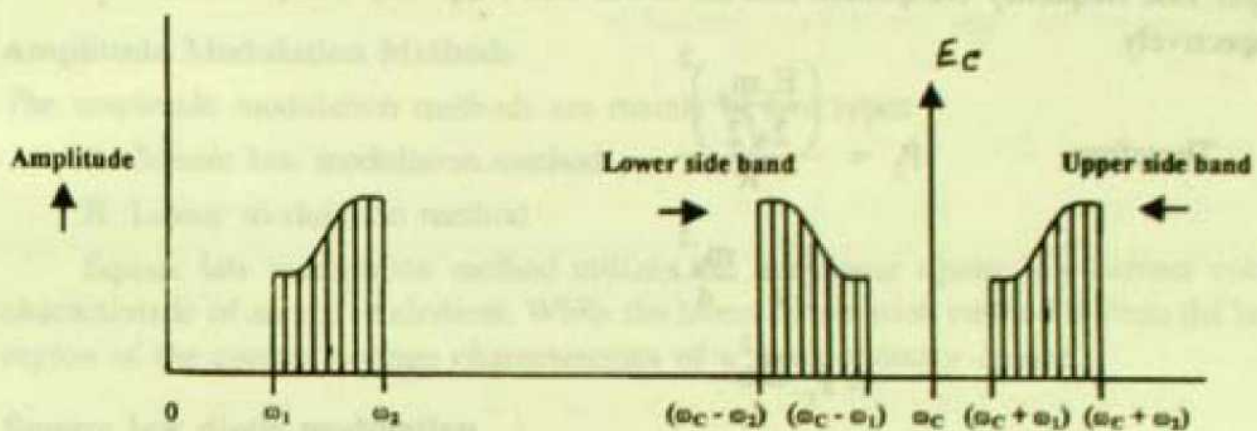


Fig.11.4 Sidebands of AM wave

The lower sideband as well as the upper sideband contain the full information while the carrier voltage  $e_c = E_c \cos \omega_c t$  contains no information. The information which was initially at a band of frequencies above zero frequency has now been shifted to two bands of frequencies one in the frequency range  $(\omega_c + \omega_1)$  to  $(\omega_c + \omega_2)$  and the other at  $(\omega_c - \omega_2)$  to  $(\omega_c - \omega_1)$  in the upper and lower side respectively. The two sidebands embody the full intelligence separately. But the total bandwidth or channel width required for the full transmission of the amplitude modulated wave is from  $(\omega_c - \omega_2)$  to  $(\omega_c + \omega_2)$ . This bandwidth is twice the highest frequency component of the modulating voltage. For the modulating voltage having frequencies 0 to 5 kHz, the amplitude modulated wave should have a channel width of 10 kHz. Broadcasting stations which have well separated carrier frequencies can broadcast amplitude modulated wave without interference though they have their individual modulating signal at the audio frequency range.

### Power in the AM wave

The amplitude modulated wave has three components

$$e = E_c \cos \omega_c t + \frac{E_c m_a}{2} \cos (\omega_c + \omega_m)t + \frac{E_c m_a}{2} \cos (\omega_c - \omega_m)t$$

The carrier frequency component of the AM wave is exactly the same as the unmodulated carrier voltage. So, the power of the carrier frequency component of the modulated wave is identical to that of the unmodulated carrier. If the power is delivered to a load resistance  $R$ , the power due to carrier frequency component of voltage is given by

$$P_c = \frac{\left(\frac{E_c}{\sqrt{2}}\right)^2}{R} = \frac{E_c^2}{2R}, \text{ where } \frac{E_c}{\sqrt{2}} \text{ is the r.m.s. value}$$

The two side frequency components of voltage have the same amplitude  $\frac{E_c m_a}{2}$ . So, they will deliver identical power to the resistance  $R$ . Let the power delivered by the upper side frequency component and the lower side frequency component be  $P_2$  and  $P_1$  respectively.

Therefore

$$\begin{aligned} P_2 &= \frac{\left(\frac{E_c m_a}{2\sqrt{2}}\right)^2}{R} \\ &= \frac{E_c^2}{2R} \cdot \frac{m_a^2}{4} \\ &= P_c \cdot \frac{m_a^2}{4} \end{aligned}$$

Power in the lower side frequency component

$$\begin{aligned} P_1 &= \frac{\left(\frac{E_c m_a}{2\sqrt{2}}\right)^2}{R} \\ &= P_c \cdot \frac{m_a^2}{4} \end{aligned}$$

The total power in the two side frequency components

$$\begin{aligned} P_s &= P_1 + P_2 \\ &= P_c \cdot \frac{m_a^2}{2} \end{aligned}$$

The total power in the amplitude modulated wave is given by,

$$\begin{aligned} P_t &= P_c + P_1 + P_2 = P_c + P_s \\ &= P_c + P_c \cdot \frac{m_a^2}{2} \\ &= P_c \left(1 + \frac{m_a^2}{2}\right) \end{aligned}$$

The power of the amplitude modulated wave is greater than the power of the unmodulated carrier. When modulation is 100%, the total power will be

$$\begin{aligned} P_t &= P_c \left(1 + \frac{1}{2}\right) \\ &= 1.5 P_c \end{aligned}$$



This shows that at 100% modulation, out of 150 units of power, 100 units are in the carrier frequency component which carries no information while 25 units are in each side frequency component which contains full information. At lower percentage modulation of AM wave, relatively lower power will be in each side frequency component.

### Amplitude Modulation Methods

The amplitude modulation methods are mainly of two types

- A. Square law modulation method
- B. Linear modulation method

Square law modulation method utilizes the non-linear square law current voltage characteristic of an active element. While the linear modulation method utilizes the linear region of the current voltage characteristics of a semiconductor device.

#### Square law diode modulation

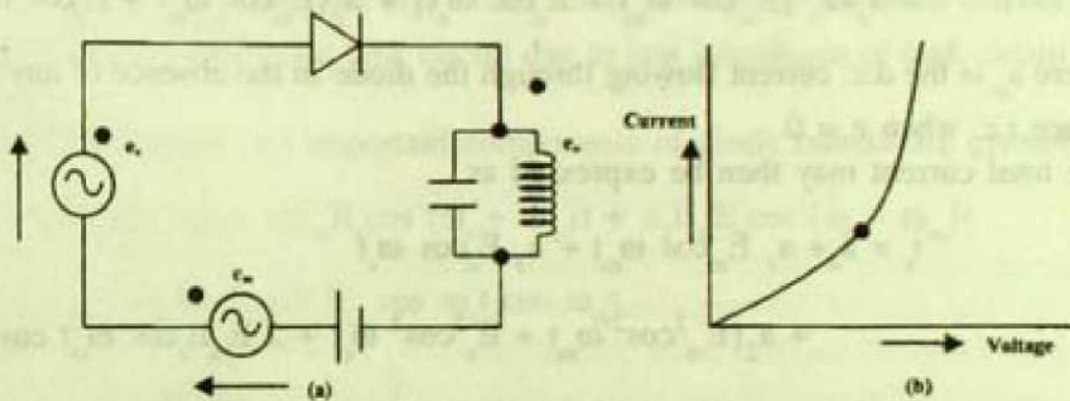


Fig. 11.5 (a) Basic circuit diagram for AM modulation using diode  
(b) dynamic current-voltage characteristics of a diode

The basic method of producing AM wave using non-linear current voltage dynamic characteristic is highly non linear as shown particularly in the low voltage operating region. The modulating voltage  $e_m$  and the carrier voltage  $e_c$  are in series with the properly biased diode along with tuned circuit (Fig. 11.5).

The a.c. diode current  $i$  may be expressed in terms of applied a.c. voltage and is given by

$$i = a_1 e + a_2 e^2$$

using upto second degree while  $a_1$  and  $a_2$  are two constants and  $e$  is the total a.c. voltage applied. But

$$e = e_m + e_c$$

Let the modulating voltage be

$$e_m = E_m \cos \omega_m t$$

and the carrier voltage be

$$e_c = E_c \cos \omega_c t$$

Therefore the total a.c. voltage applied is

$$\begin{aligned} e &= e_m + e_c \\ &= E_m \cos \omega_m t + E_c \cos \omega_c t \end{aligned}$$

The total current  $i_t$  is due to the d.c. biasing voltage and the applied a.c. voltage and is given by,

$$\begin{aligned} i_t &= a_0 + a_1 e + a_2 e^2 \\ &= a_0 + a_1 \{E_m \cos \omega_m t + E_c \cos \omega_c t\} + a_2 \{E_m \cos \omega_m t + E_c \cos \omega_c t\}^2 \end{aligned}$$

where  $a_0$  is the d.c. current flowing through the diode in the absence of any applied a.c. voltage i.e. when  $e = 0$ .

The total current may then be expressed as

$$\begin{aligned} i_t &= a_0 + a_1 E_m \cos \omega_m t + a_1 E_c \cos \omega_c t \\ &\quad + a_2 \{E_m^2 \cos^2 \omega_m t + E_c^2 \cos^2 \omega_c t + 2 E_m E_c \cos \omega_m t \cos \omega_c t\} \\ &= a_0 + a_1 E_m \cos \omega_m t + a_1 E_c \cos \omega_c t \\ &\quad + a_2 E_m^2 \left( \frac{1 + \cos 2\omega_m t}{2} \right) + a_2 E_c^2 \left( \frac{1 + \cos 2\omega_c t}{2} \right) \\ &\quad + a_2 E_m E_c \{ \cos (\omega_c + \omega_m) t + \cos (\omega_c - \omega_m) t \} \end{aligned}$$

The various frequency components of the current may then be identified. These components are (in order of frequencies)

d.c. component:  $a_0$ ,  $\frac{a_2 E_m^2}{2}$  and  $\frac{a_2 E_c^2}{2}$

modulating frequency component:  $a_1 E_m \cos \omega_m t$

second harmonic of  $\omega_m$  component:  $\frac{a_2 E_m^2}{2} \cos 2 \omega_m t$

$(\omega_c - \omega_m)$  frequency component:  $a_2 E_m E_c \cos (\omega_c - \omega_m) t$



$\omega_c$  carrier frequency component:  $a_1 E_c \cos \omega_c t$

$\omega_c + \omega_m$  frequency component:  $a_2 E_m E_c \cos (\omega_c + \omega_m) t$

second harmonic of  $\omega_c$  component:  $\frac{a_2 E_c^2}{2} \cos 2 \omega_c t$

In addition to the carrier frequency term  $a_1 E_c \cos \omega_c t$ , the term  $a_2 E_m E_c \cos (\omega_c - \omega_m) t$  represents the lower side frequency and the term  $a_2 E_m E_c \cos (\omega_c + \omega_m) t$  represents the upper side frequency. The load impedance as shown is a tuned circuit which is tuned at the carrier frequency  $\omega_c$ . It responds quite well to a narrow band of frequency centered around the carrier frequency  $\omega_c$ . Hence the frequency component which primarily contribute to develop the output are  $\omega_c$ ,  $\omega_c - \omega_m$  and  $\omega_c + \omega_m$ . This is true when upper side frequency and lower side frequency are close to  $\omega_c$  i.e.  $\omega_c \gg \omega_m$ . The rest of the terms will not produce considerable voltage across the tank circuit due to low impedance of tank circuit at these frequencies.

Hence the desired and important components of diode current are given by,

$$i_d = a_1 E_c \cos \omega_c t + a_2 E_m E_c \cos (\omega_c + \omega_m) t + a_2 E_m E_c \cos (\omega_c - \omega_m) t$$

$$= a_1 E_c \cos \omega_c t + 2 a_2 E_c E_m \cos \omega_c t \cos \omega_m t$$

$$= a_1 E_c \left\{ 1 + \frac{2a_2 E_m}{a_1} \cos \omega_m t \right\} \cos \omega_c t$$

$$= a_1 E_c \{ 1 + m_a \cos \omega_m t \} \cos \omega_c t$$

where  $m_a = \frac{2a_2 E_m}{a_1}$ , is the modulation index. The modulated output voltage across

the tank circuit may be expressed by,

$$e_o = i_d R_t$$

$$= R_t a_1 E_c (1 + m_a \cos \omega_m t) \cos \omega_c t$$

where  $R_t$  is the impedance of the tank circuit at and around the carrier frequency  $\omega_c$ .

### Linear modulation method

The following is the circuit diagram of an amplitude modulated amplifier with emitter modulation. Here the modulating voltage is applied to the emitter and AM wave is obtained at the output.

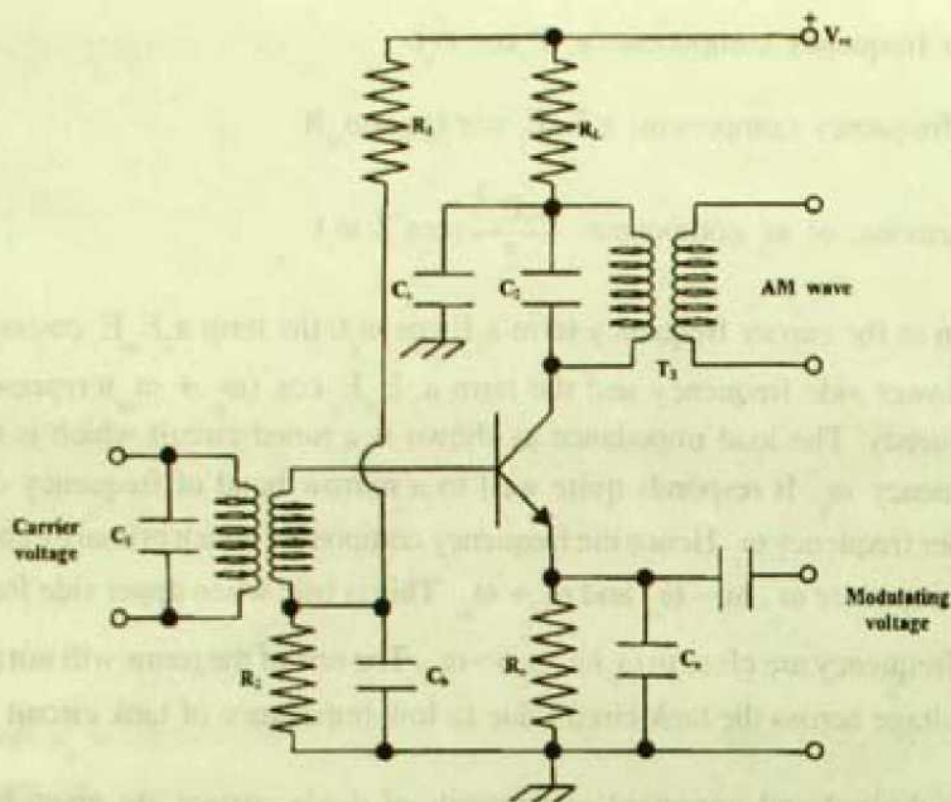


Fig. 11.6 Amplitude modulating amplifier

The modulating voltage and the carrier voltage are applied to the same amplifier (Fig. 11.6). The gain of the amplifier is varied by the modulating signal and the output of the amplifier gives the AM wave. This amplifier is called the modulating amplifier or simply modulator. The modulating signal is applied to the emitter and the transistor amplifier operates in the CE mode. The carrier voltage is applied to the base of the transistor through the input transformer. Resistances  $R_1$ ,  $R_2$ ,  $R_c$ , and  $R_L$  along with the supply voltage establish the quiescent operating point of the CE amplifier.  $C_b$ ,  $C_c$ , and  $C_e$  are by pass capacitors used to by pass carrier frequency voltage.  $C_1$  and  $C_2$  are two capacitors used to tune the primary of the single tuned transformer. Tuned circuit  $T_2$  has the sufficient bandwidth to accommodate the carrier frequency, the upper side frequency and the lower side frequency of the AM signal obtained at the output. The modulating voltage applied to the emitter changes the emitter current from its quiescent value and the variation is proportional to the modulating voltage. Therefore the instantaneous emitter current may be expressed by,

$$i_E = I_E + K_1 E_m \cos \omega_m t$$

Where  $I_E$  = quiescent emitter current

$K_1$  = a proportionality constant

and  $E_m \cos \omega_m t$  = modulating voltage

$\omega_m$  = modulating frequency



As already mentioned the voltage amplification  $A_v$  is controlled by the instantaneous value of the emitter current,  $i_E$ . Therefore,

$$\begin{aligned} A_v &= K_2 i_E \\ &= K_2 [I_E + K_1 E_m \cos \omega_m t] \end{aligned}$$

where  $K_2$  is another proportionality constant.

Let the carrier voltage which is applied at the base of the amplifier is

$$e_c = E_c \cos \omega_c t$$

This being the input voltage of the amplifier, the output voltage will be given by,

$$\begin{aligned} e_o &= A_v \cdot e_c \\ &= K_2 [I_E + K_1 E_m \cos \omega_m t] E_c \cos \omega_c t \\ &= K_2 E_c I_E \left[ 1 + \frac{K_1}{I_E} E_m \cos \omega_m t \right] \cos \omega_c t \\ &= K [1 + m_a \cos \omega_m t] \cos \omega_c t \end{aligned}$$

where  $m_a$  = depth of modulation =  $\frac{K_1 E_m}{I_E}$  and  $K = K_2 E_c I_E$ .

The output voltage  $e_o$  has the nature of AM signal. The amplitude of the modulated voltage varies with the instantaneous values of the modulating voltage  $E_m \cos \omega_m t$ . As usual, AM signal will have three parts one each at  $\omega_c$ ,  $\omega_c + \omega_m$  and  $\omega_c - \omega_m$  rotational frequency. The three components may be expressed by

$$K \cos \omega_c t \text{ (carrier frequency component)}$$

$$K \frac{m_a}{2} \cos (\omega_c + \omega_m) t \text{ (upper side frequency component)}$$

and  $K \frac{m_a}{2} \cos (\omega_c - \omega_m) t$  (lower side frequency component)

The symbol  $K$  contains the amplitude of the carrier voltage,  $E_c$ .

### Limitation of AM

From the discussion so far it may appear that amplitude modulation is highly effective in sending low frequency modulating signal. But practically there are a number of limitations in it.

#### A. Low power carrying capacity :

It has been mentioned that the information or intelligence remains in the side frequencies and no information is in the carrier frequency component of the modulated signal. Even when the percentage modulation is 100%, the power contained in the two side frequencies or side bands, as the case may be, is only one third of the total power transmitted. As the information remains in the side bands, it is apparent that the efficiency in carrying information through amplitude modulation is very poor.

#### B. Interference by noise:

In the amplitude modulation process, the amplitude of the carrier voltage is varied in accordance with the signal or information or intelligence. Almost all man-made noise and natural noise are of different amplitudes. A radio receiver cannot distinguish the sources which produce the variation of amplitude of the received wave in other words it will receive the unwanted noise and the wanted amplitude modulated wave simultaneously. Thus the reception of amplitude modulated wave is more noisy.

#### C. Poor audio quality:

For quality audio reception, frequencies upto 15 kHz have to be reproduced at the receiving end quite well. This requires a 30 kHz bandwidth to accommodate both the upper and the lower sidebands. But amplitude modulated broadcasting is permitted to have a bandwidth of only 10 kHz. This is invariably done to accommodate a large number of stations. With this regulation, the highest audio frequency that can be used is merely 5 kHz. This frequency limitation cuts down the quality of the music or even speech.

### 11.3 Frequency Modulation

Frequency modulation is the process of varying the frequency of the carrier voltage in accordance with the instantaneous value of the modulating voltage. The amplitude of the carrier voltage does not change due to frequency modulation i.e. the amplitude of the modulated wave remains always unaltered.

#### Expression for FM wave

Let the carrier voltage be represented by,

$$\begin{aligned} e_c &= E_c \cos (\omega_c t + \theta) \text{ where } \theta \text{ is initial phase angle} \\ &= E_c \cos \phi \end{aligned}$$

where  $\phi$  = total instantaneous phase angle of the carrier

$$= \omega_c t + \theta$$

The angular frequency  $\omega_c$  of the carrier voltage is related to the total phase angle  $\phi$  by

$$\frac{d\phi}{dt} = \omega_c$$



In frequency modulation (Fig. 11.7), the frequency of the carrier voltage no longer remains constant but varies with time in accordance with the instantaneous value of the modulating signal. Therefore the frequency of the modulated wave after frequency modulation be given by,

$$\omega = \omega_c + \text{variation of frequency.}$$

The variation of frequency is proportional to the modulating voltage. Let the modulating voltage be

$$e_m = E_m \cos \omega_m t$$

Therefore, instantaneous frequency of the modulated wave will be

$$\omega = \omega_c + K_f E_m \cos \omega_m t$$

where  $K_f$  is the proportionality constant, representing frequency conversion from the modulating voltage.

Integration of the above equation gives the phase angle of the modulated wave. On integration

$$\begin{aligned} \phi &= \int \omega dt \\ &= \int (\omega_c + K_f E_m \cos \omega_m t) dt \\ &= \omega_c t + \frac{K_f E_m}{\omega_m} \sin \omega_m t + \theta_1 \end{aligned}$$

where  $\theta_1$  is the constant of integration.  $\theta_1$  is a constant phase angle and may be neglected in the analysis as it is insignificant in the modulation process.

The modulated voltage is therefore

$$\begin{aligned} e &= E_c \cos \phi \\ &= E_c \cos \left[ \omega_c t + \frac{K_f E_m}{\omega_m} \sin \omega_m t \right] \end{aligned}$$

Instantaneous frequency of the modulated wave

$$\begin{aligned} f &= \frac{\omega}{2\pi} = \frac{1}{2\pi} (\omega_c + K_f E_m \cos \omega_m t) \\ &= f_c + \frac{K_f E_m}{2\pi} \cos \omega_m t \end{aligned}$$

The value of  $\cos \omega_m t$  varies from 1 to -1

Therefore the maximum frequency of the modulated wave

$$f_{\max} = f_c + \frac{K_f E_m}{2\pi}$$

and

$$f_{\min} = f_c - \frac{K_f E_m}{2\pi}$$

The frequency deviation i.e. the maximum variation of frequency from the carrier frequency  $f_c$  is

$$\Delta f = \frac{K_f E_m}{2\pi} = f_{\max} - f_c = f_c - f_{\min}$$

The frequency deviation  $\Delta f$  is proportional to  $E_m$ , the amplitude of the modulating voltage but independent of the frequency of the modulating voltage.

The ratio of the maximum frequency deviation  $\Delta f$  to the modulation frequency  $f_m (= \frac{\omega_m}{2\pi})$  is known as the modulation index for FM wave and generally denoted by  $m_f$

$$\therefore m_f = \frac{\Delta f}{f_m} = \frac{K_f E_m}{2\pi} \cdot \frac{1}{f_m} = \frac{K_f E_m}{\omega_m}$$

Therefore the frequency modulated wave may be expressed as

$$\begin{aligned} e &= E_c \cos \left[ \omega_c t + \frac{K_f E_m}{\omega_m} \sin \omega_m t \right] \\ &= E_c \cos \left[ \omega_c t + m_f \sin \omega_m t \right] \end{aligned}$$

where  $m_f$  is the frequency modulation index.

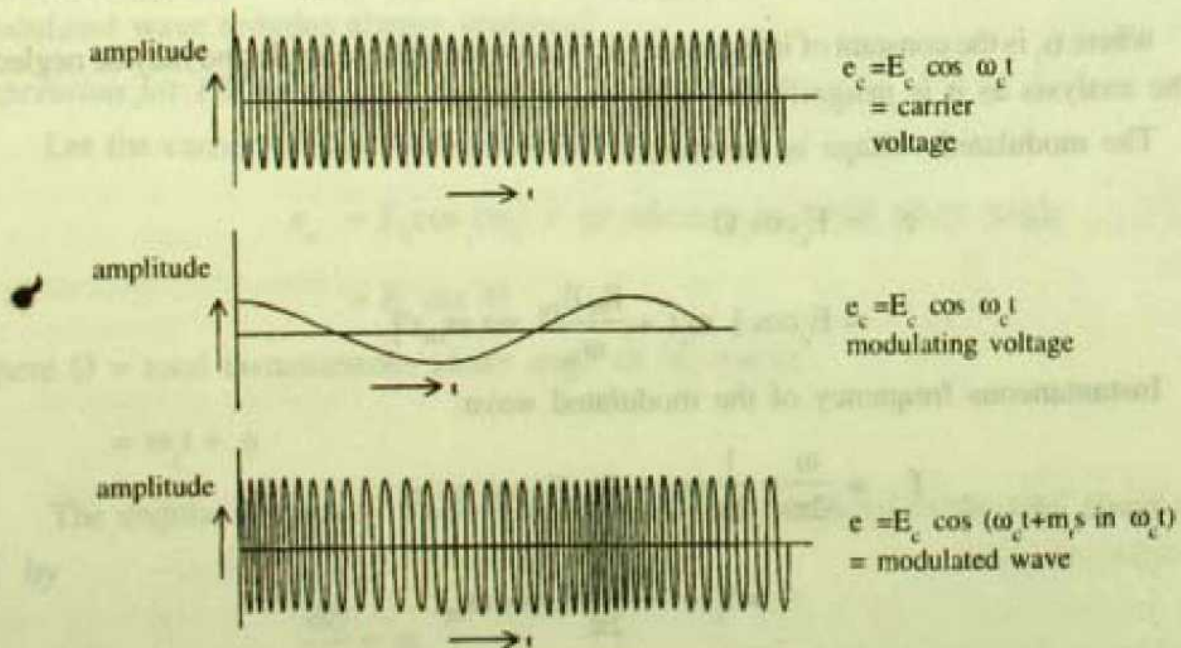


Fig. 11.7 Waveform of FM wave



When a carrier voltage is frequency modulated by some modulating voltage the amplitude of the modulated voltage remains unaltered but its frequency is varied in accordance with the time variation of the modulating voltage as shown in Figure 11.7.

### Frequency Spectrum of FM wave

The mathematical analysis of the expression of FM wave will lead to the spectrum of frequency modulated wave but this method of analysis is very complicated than that for the AM wave. The results of such analysis may only be discussed here. The frequency spectrum of FM wave consists of a carrier component together with side frequencies which are harmonics of the modulating frequency. It means that for a modulating voltage of frequency  $w_m$  the frequency spectrum of the FM wave may contain components of frequency  $w_c \pm nw_m$ . The components having different amplitudes at different frequencies add together and give a constant amplitude for a FM wave. The amplitudes of the various components of the spectrum are given by a mathematical function known as the Bessel function of the first kind. For a particular value of frequency modulation index,  $m_f$ , this function is usually denoted by  $J_n(m_f)$  where  $n$  is the order of the side frequency. Bessel functions are available both in graphical and tabular form (Table 11.1).  $J_0(m_f)$  gives the amplitude of carrier component for modulation index  $m_f$ . Amplitudes of spectrum component are for unmodulated amplitude 1 volt. Amplitude values having 1% of unmodulated amplitude were neglected.

**Table 11.1 Bessel Function Amplitude**

Modulation index, $m_f$	carrier $J_0$	Side frequency							
		$J_1$	$J_2$	$J_3$	$J_4$	$J_5$	$J_6$	$J_7$	$J_8$
.25	.98	.12							
.50	.94	.24	.03						
1.0	.77	.44	.11	.02					
2.0	.22	.58	.35	.13	.03				
(2.4)	0	.52	.43	.20	.06				
4.0	-.40	-.07	.36	.43	.28	.13	.05	.02	

As an example of the use of the table, it can be seen that for  $m_f = .5$ , the spectral components are

carrier ( $f_c$ ) [usually called centre frequency]  $J_{0(.5)} = .94$

first order side frequencies ( $f_c \pm f_m$ )  $J_{1(.5)} = .24$

second order side frequencies ( $f_c \pm 2f_m$ )  $J_{2(.5)} = .03$



The fact that the spectrum component at the carrier (or center) frequency decreases in amplitude does not indicate that it is amplitude modulated. The modulated wave is the sum of all the spectrum components which are all sine waves. All spectrum components are either sine or cosine waves. Their amplitudes may be positive or negative but usually it is not necessary to show then on spectrum graphs. For certain values of  $m_f$  (2.4, 5.5, 8.65 etc), the carrier amplitude goes to zero. This points out that it is the sinusoidal component of the spectrum at carrier frequency which goes to zero amplitude but not the modulated voltage. The spectrum of FM wave for two values of modulation index,  $m_f = 1.0$  and  $m_f = 2.4$  are shown in (Fig 11.8).

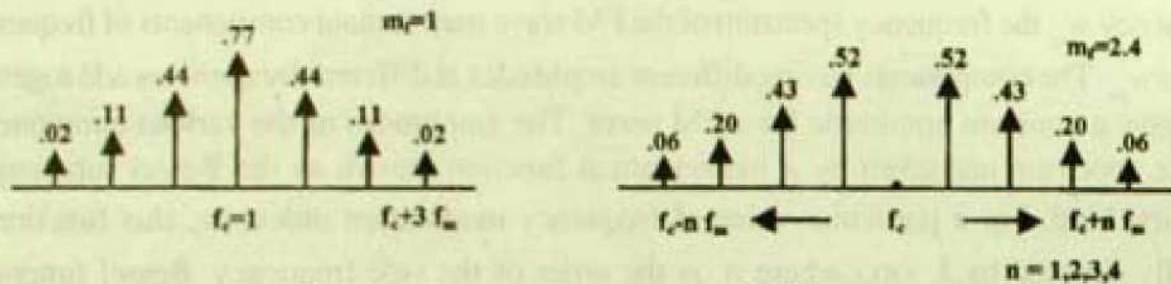


Fig.11.8 Frequency spectrum of FM wave

### Frequency Modulation Method

If the capacitance or inductance of an LC oscillator circuit is made to vary the frequency of oscillation of the circuit varies. If these variations are made directly proportional to the modulating voltage true frequency modulated wave may be obtained. If such a variable reactance circuit is placed across the tank circuit of an LC oscillator, the effective frequency of the oscillator can be made to vary in accordance with the applied voltage. The larger the variation of the modulating voltage from its zero voltage level, the larger will be the variation of the frequency of oscillation of the tank circuit, thus producing true frequency modulation. When the modulating voltage is zero the variable capacitance or inductance will produce its mean or average value in shunt with the tank circuit reactive elements. Thus the resultant carrier or centre frequency is determined by the LC tank circuit along with the added average inductance or capacitance offered by the modulator circuit.

### Basic Reactance Modulator

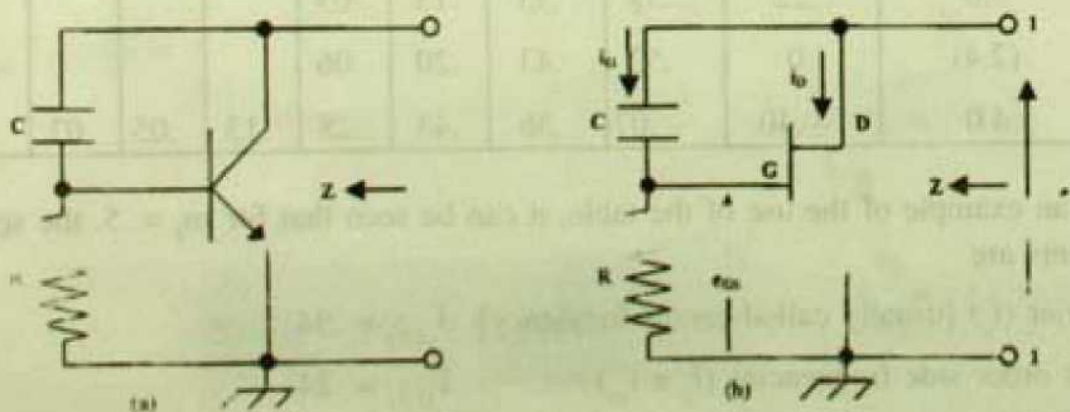


Fig. 11.9 (a) modulation using transistor (b) modulation using FET



The above two basic circuits (Fig. 11.9) one using transistor and the other using FET can be used to explain the working principle of the reactance modulator. Let the impedance looking left from the 1,1 terminal be represented by  $Z$ . The modulator reactance appears between the drain and source of the FET modulator i.e. between 1,1 terminals as shown, while its value may be controlled by the modulating voltage applied between the gate and the source. The source  $S$  being common to the input and output, it is generally called a three terminal reactance. This three terminal reactance is usually connected across the tank circuit of an oscillator to produce FM wave.

### Theory of the reactance modulator

Let some applied voltage between 1,1 be  $e$ . Due to this voltage drain current  $i_D$  will flow through the FET.

Let the gate current be  $i_G$ .

The capacitance  $C$  and resistance  $R$  are so chosen that at the frequency concerned

$$\frac{1}{\omega C} \gg R$$

As the input impedance of the FET is very large than  $R$  used in the circuit,  $i_G$  may be expressed as

$$i_G = \frac{e}{R + \frac{1}{j\omega C}}$$

The gate to source voltage

$$e_{GS} = i_G \cdot R$$

$$= \frac{eR}{R + \frac{1}{j\omega C}}$$

$$= \frac{eR}{\frac{1}{j\omega C}} \quad \text{when } \frac{1}{\omega C} \gg R$$

$$= e \cdot j\omega CR$$

The FET drain current

$$i_D = g_m \cdot e_{GS}$$

where  $g_m$  = transconductance of the FET used

$$= e \cdot j\omega CR g_m$$

Therefore the impedance  $Z$  seen from 1,1 terminals due to the reactance modulator is

$$\begin{aligned}
 Z &= \frac{e}{i_D} \\
 &= \frac{e}{e \cdot j\omega CR g_m} \\
 &= \frac{1}{j\omega CR g_m} \\
 &= \frac{1}{j\omega C_{eq}}
 \end{aligned}$$

where  $C_{eq} = CRg_m$ , is the equivalent capacitance offered by the modulator between 1,1 terminal. Modulating voltage when applied between the gate and the source of the FET, varies the  $g_m$  of the FET, thereby producing a variation of the equivalent capacitance,  $C_{eq}$ . This capacitance  $C_{eq}$  when placed across the tank circuit of an oscillator produces perfect

FM wave. It may be noted here that this perfect FM is obtained by assuming  $\frac{1}{\omega C} \gg R$

and neglecting  $R$  in comparison to  $\frac{1}{\omega C}$ . i.e  $R + \frac{1}{j\omega C} = \frac{1}{j\omega C}$

If  $R$  cannot be neglected totally, then

$$e_{GS} = \frac{e \cdot j\omega CR}{e \cdot j\omega CR}$$

$$i_D = \frac{e \cdot j\omega CR g_m}{1 + j\omega CR}$$

and

$$\begin{aligned}
 Z &= \frac{1 + j\omega CR}{j\omega CR g_m} \\
 &= \frac{1}{j\omega CR g_m} + \frac{1}{g_m} \\
 &= \frac{1}{j\omega C_{eq}} + \frac{1}{g_m}
 \end{aligned}$$

The reactive part produces the required frequency modulation. But the resistive part of the impedance damps the tank circuit producing small amount of amplitude variation. This amplitude variation once again depends on the modulating voltage because  $g_m$  varies in accordance with the modulating voltage.



If the reactance modulator circuit is modified as Fig. 11.10.

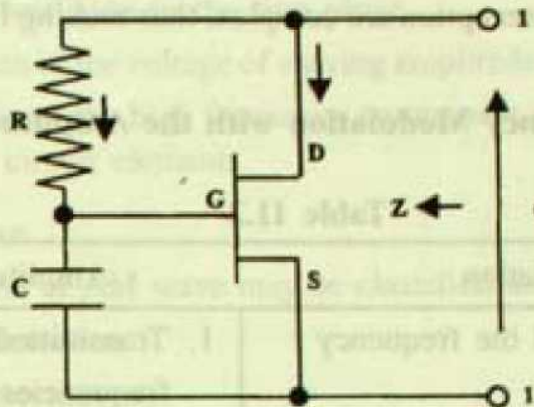


Fig. 11.10 Reactance modulator circuit

and if  $R \gg \frac{1}{\omega C}$ , it can be shown that the equivalent reactance is that of an inductance

$L_{eq}$  where,

$$L_{eq} = \frac{RC}{g_m}$$

To make the idea of the above principle clear, a complete circuit diagram using a transistor reactance modulator is given in Fig. 11.11.

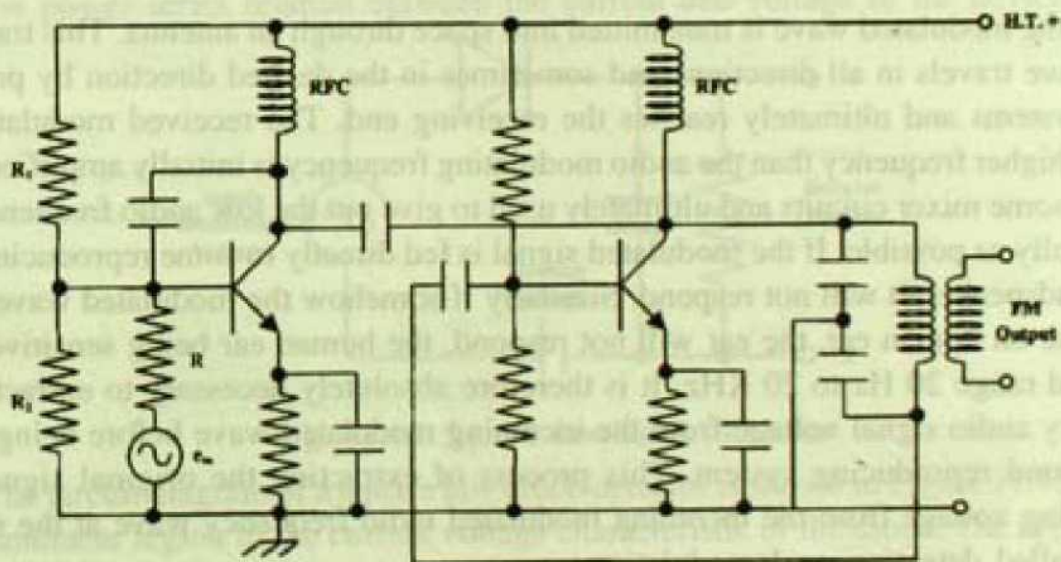


Fig. 11.11 Transistor reactance modulator using Colpitt's oscillator

### Few facts about Frequency Modulation

In all television transmissions audio signal is transmitted through frequency modulation. In the mobile radio services and also in the amateur bands audio frequencies ranging from 20 Hz to 4000 Hz are used along with frequency modulation. FM entertainment broadcasting is made in the frequency band 88 to 108 MHz with a channel width of 200 kHz. The



maximum frequency deviation is  $\pm 75$  KHz. The circuitry and equipments used for frequency modulation transmission and reception are complex, thus making FM more expensive than AM.

### Comparison of the Frequency Modulation with the Amplitude Modulation

Table 11.2

Frequency modulation	Amplitude modulation
1. Transmitted powers in all the frequency components are useful	1. Transmitted powers in the side frequencies are useful
2. Power required for transmission is high	2. Power required for transmission is low
3. Transmitted power is less sensitive to noise	3. Transmitted power is highly sensitive to noise
4. Signal to noise ratio is high	4. Signal to noise ratio is low
5. Frequency modulation index value may be larger than unity	5. modulation index is always less than unity

### 11.4 Detection or Demodulation

Usually the modulated wave is transmitted into space through an antenna. This transmitted radio wave travels in all directions and sometimes in the desired direction by proper use of the systems and ultimately reaches the receiving end. The received modulated wave at much higher frequency than the audio modulating frequency is initially amplified, passed through some mixer circuits and ultimately used to give out the low audio frequency signal as faithfully as possible. If the modulated signal is fed directly to some reproducing device say a loudspeaker, it will not respond. Similarly if somehow the modulated wave is made to operate on human ear, the ear will not respond, the human ear being sensitive only in the broad range 20 Hz to 20 KHz. It is therefore absolutely necessary to extract the low frequency audio signal voltage from the incoming modulated wave before being fed into some sound reproducing system. This process of extracting the original signal or the modulating voltage from the incoming modulated radio frequency wave at the receiving end is called detection or demodulation

The process of demodulation or detection of amplitude modulated radio waves consists of two parts —

- rectification of the amplitude modulated wave and
- elimination of the high frequency component of the AM wave.

But the demodulation or detection of frequency modulated wave consists of three parts in succession —



- (a) conversion of the frequency variation of the FM wave into corresponding amplitude variation in the first phase
- (b) rectification of the voltage of varying amplitude obtained in the first phase and
- (c) elimination of the high frequency component of the modulated wave by use of proper circuit elements.

### Detection of AM wave

The process of detection of AM wave may be classified into the following two groups—

1. Linear diode detection
2. Square law detection

The rectification property of a diode is essential for detection. A linear detector uses the property of linear rectification. In linear diode detection, the rectification property of detector diode is used in such a way that current through the diode flows in pulses. The behavior of the device cannot, therefore, be analysed by power series and in this case a linear relation exists between the amplitude of the AM wave and that of the detected output voltage.

On the other hand, the square law detection uses the non linear relation usually the square law relation between the current voltage characteristics of a device. The active device may be a diode or a transistor. In the square law detector, the current flowing through the device is continuous and hence the behaviour of this circuit can be analysed properly by using the power series relation between the current and voltage of the device.

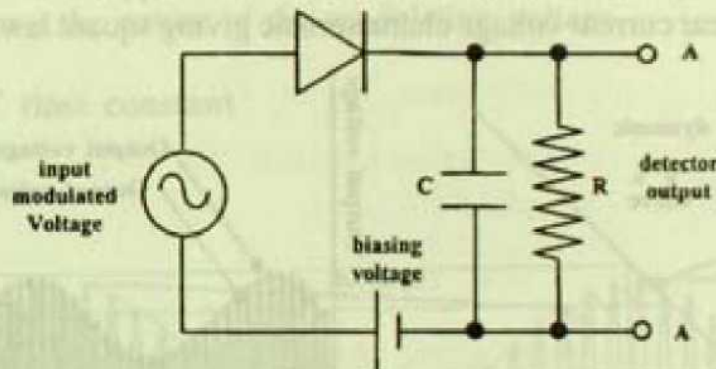


Fig. 11.12 Square law diode detector

The circuit diagram of a square law diode detector is shown in Figure 11.12. It works in the nonlinear region of the current voltage characteristic of the diode. The applied input modulated voltage is of small amplitude. Whereas in the linear diode detector, the applied modulated voltage is large in magnitude and the operation is in the linear region of the characteristic. In the square law diode detector, the diode is biased positively to fix the operating point to a small current non linear region of the dynamic characteristics of the current and voltage. The resistance capacitance (RC) combination constitute the load impedance of the detector across which the detected output is obtained. As a matter of fact, the resistance R alone is responsible for square law detection in combination with the applied biasing voltage and input modulated voltage. The capacitance C performs the



function of by passing the high frequency component of voltage dropped across R in absence of the capacitor C.

### Linear Diode Detector

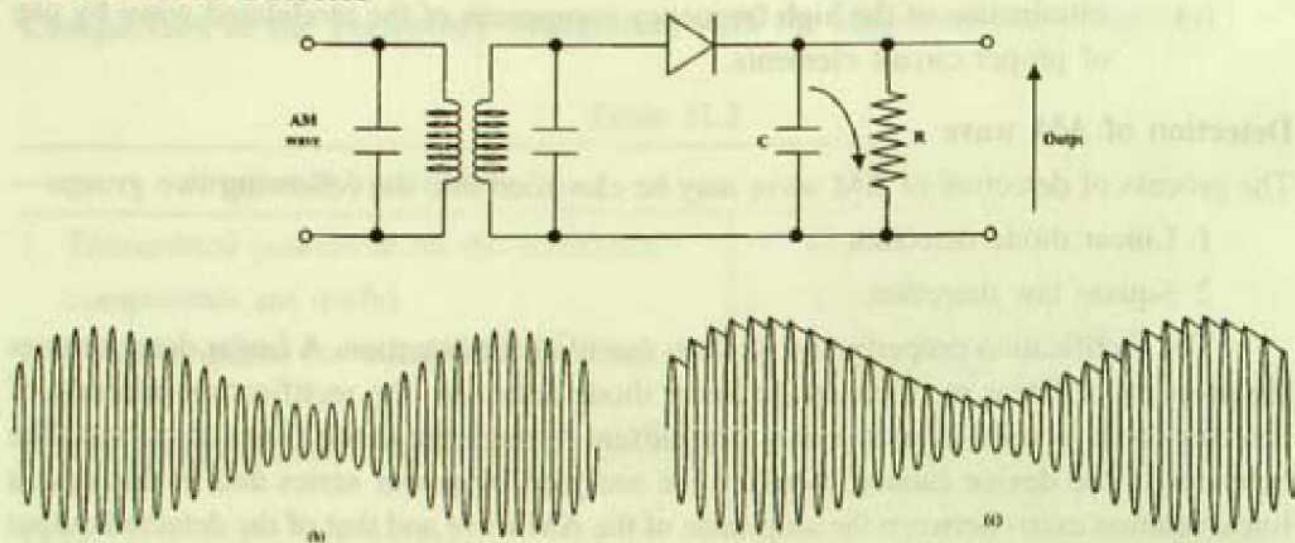


Fig. 11.13 (a) Linear diode detector (b) input AM wave (c) detector output

The diagrams (Fig. 11.13) show the circuit of linear diode detector as well as the input AM wave and detected output voltage. The linear diode detector is extensively used in commercial radio receivers. For satisfactory operation, the linear diode detector requires modulated carrier voltage of 1 volt or more. If the diode operates on a smaller voltage, it works in the non-linear current voltage characteristic giving square law detection response.

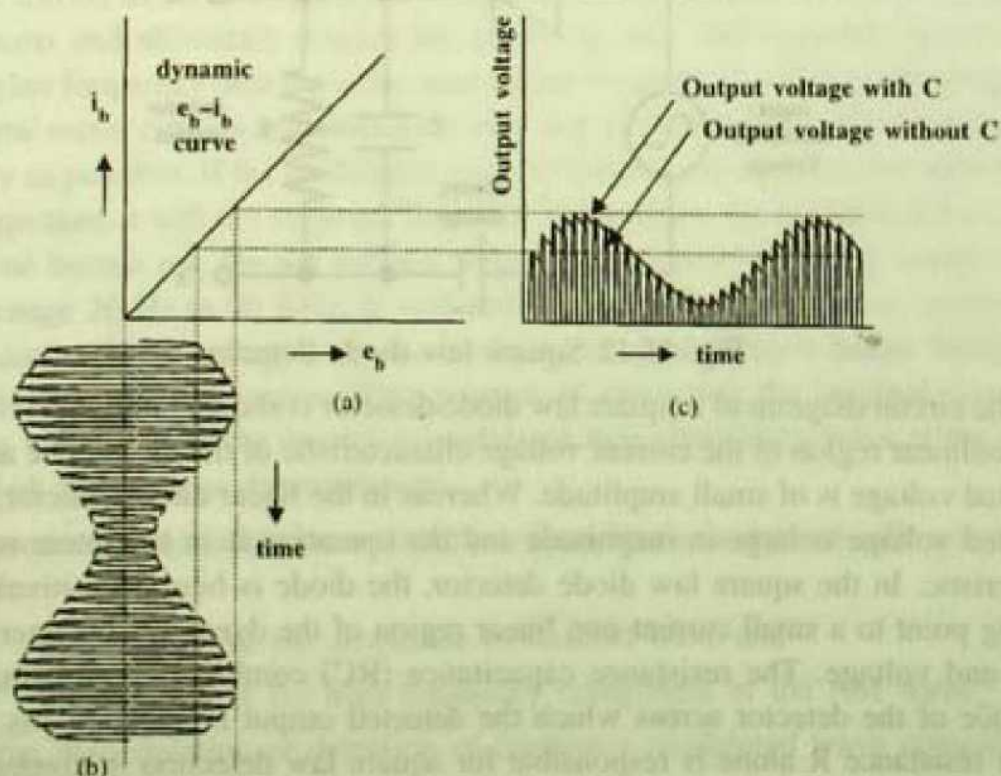


Fig.11.14 (a) Current-voltage curve (b) modulated input voltage (c) output voltage



The dynamic current voltage characteristic of the detector diode operating at a modulating voltage above one volt is shown in Fig. 11.14 along with the output voltage waveform across the load resistance  $R$  in presence of the capacitor and also in its ( $C$ 's) absence. At the primary of the transformer is the voltage which is amplitude modulated in nature, the secondary of the transformer will show the same waveform i.e. AM wave. But due to the series addition of the detector diode and load resistance  $R$ , current will flow only in the positive half cycle of the input modulated voltage in the absence of  $C$ , but not in the negative half cycle of the input modulated voltage, giving an output current pulse of half cycles. The magnitude of this current pulse will vary according to the variation of amplitude with time of the amplitude modulated wave. Hence the half-cycle current pulses are of unequal amplitude and it varies with time. As soon as the capacitor  $C$  is placed across the resistance  $R$  the nature of the output voltage which was current multiplied by resistance in the absence of  $C$ , changes dramatically. In presence of the capacitor, the voltage drop across the resistance  $R$  or the capacitor  $C$  will always be identical. There is a continuous current flow through the resistance  $R$ . When the detector diode is conducting for a small angle, the current is supplied via the transformer but when the detector diode is not conducting the current through the resistance  $R$  is supplied by the charged capacitor  $C$ . Thus the condenser  $C$  discharges and the voltage across  $C$  decreases with a time constant  $RC$ . If the rate of fall of the voltage across  $C$  is smaller than that of the envelope of the modulated voltage, there will be the diagonal clipping. With proper value of the  $RC$  product, the output voltage may be made to follow the envelope of the modulated wave. The ultimate output voltage after filtering follows the nature of the modulating voltage.

### Choice of the $RC$ time constant

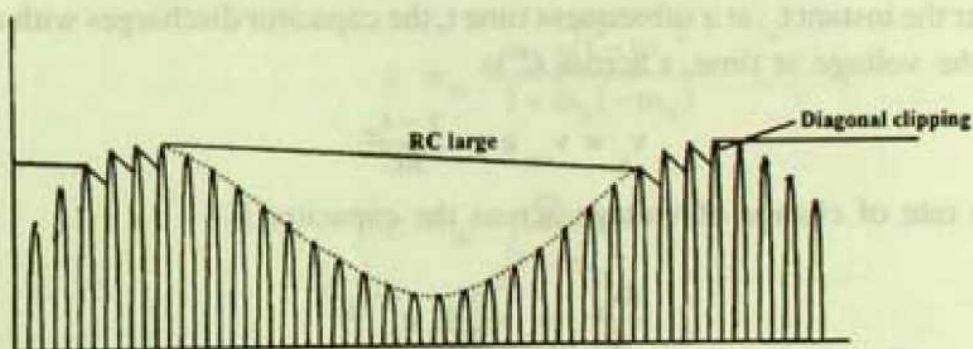


Fig.11.15 Effect of large  $RC$

As mentioned above for very large value of  $RC$  the output voltage across  $R$  cannot follow the envelope of the modulated voltage. When  $RC$  is infinite, there will be total clipping of the declining wave nature and the output will be of maximum value of the modulated voltage (Fig. 11.15). On the contrary, if the time constant is zero the output current and voltage waveform across  $R$ , will be of the nature of half sinusoids having different amplitudes. It is nothing but simple rectification in the absence of  $C$  i.e. for  $RC = 0$ ,  $C$  must be zero. It offers very low output detected voltage and the detection efficiency becomes very poor.



For the best operation the RC time constant will have a proper value for the given value of modulation index  $m_a$  and the maximum value of the modulating frequency. This is due to the fact that at the maximum modulating frequency, the waveform of the modulated voltage falls most rapidly thus creating the worst situation for diagonal clipping.

Let the equation of the envelope of the modulated voltage be,

$$v = E_c [1 + m_a \cos \omega_m t]$$

where  $\omega_m$  here is the maximum modulating frequency. Therefore the slope of the envelope is

$$\frac{dv}{dt} = -E_c \omega_m m_a \sin \omega_m t$$

The value of the voltage of the envelope at any time  $t = t_o$  is

$$v_o = E_c [1 + m_a \cos \omega_m t_o]$$

At that time  $t = t_o$ , the slope of the envelope is

$$\left(\frac{dv}{dt}\right)_{t=t_o} = -E_c \omega_m m_a \sin \omega_m t_o$$

Let  $t_o$  be the instant when the capacitor starts discharging through the resistance R and just at this time the voltage across the capacitor must be equal to the envelope voltage. Let this voltage be represented by  $v_{co}$ .

$$\begin{aligned} \therefore v_{co} &= v_o \\ &= E_c [1 + m_a \cos \omega_m t_o] \end{aligned}$$

After the instant  $t_o$ , at a subsequent time  $t$ , the capacitor discharges with a time constant RC and the voltage at time,  $t$  across C is

$$v_c = v_{co} e^{-\frac{t-t_o}{RC}}$$

The rate of change of voltage across the capacitor is

$$\frac{dv_c}{dt} = -\frac{v_{co}}{RC} e^{-\frac{t-t_o}{RC}}$$

At time  $t = t_o$ , the rate of change of voltage across the capacitor is given by

$$\begin{aligned} \left(\frac{dv}{dt}\right)_{t=t_o} &= -\frac{v_{co}}{RC} \\ &= -\frac{1}{RC} \cdot E_c [1 + m_a \cos \omega_m t_o] \end{aligned}$$

To avoid diagonal clipping, the slope of the capacitor voltage  $v_c$  at  $t = t_o$  should be algebraically equal to or less than the slope of the envelope of the modulated voltage.



$$\text{So, } -\frac{1}{RC} \cdot E_c [1+m_a \cos \omega_m t_o] \leq -E_c \omega_m m_a \sin \omega_m t_o$$

$$\text{or, } \frac{1}{RC} E_c [1+m_a \cos \omega_m t_o] \geq E_c \omega_m m_a \sin \omega_m t_o$$

The above expression justifies the condition that the rate of decay of capacitor voltage should be equal to or greater than the rate of decay of the envelope voltage.

$$\text{or } \frac{1}{RC} \geq \omega_m \cdot \frac{m_a \sin \omega_m t_o}{1+m_a \cos \omega_m t_o}$$

The worst condition is faced when  $\frac{m_a \sin \omega_m t_o}{1+m_a \cos \omega_m t_o}$  is a maximum. The condition of maximum value of this factor can be obtained by differentiating it with respect to time and putting the value to zero.

$$\frac{d}{dt} \left[ \frac{m_a \sin \omega_m t_o}{1+m_a \cos \omega_m t_o} \right] = 0 \text{ at } t = t_o$$

$$\text{It gives } \cos \omega_m t_o = -m_a$$

$$\text{and } \sin \omega_m t_o = \sqrt{1-m_a^2}$$

Thus the necessary condition is

$$\frac{1}{RC} \geq \omega_m m_a \frac{\sin \omega_m t_o}{1+\cos \omega_m t_o}$$

$$\geq \omega_m \cdot \frac{m_a \sqrt{1-m_a^2}}{1+m_a(-m_a)}$$

$$\geq \omega_m \frac{m_a}{\sqrt{1-m_a^2}}$$

With the increasing value of the modulation index,  $m_a$ , the ratio  $\frac{m_a}{\sqrt{1-m_a^2}}$  increases which restricts the value of the time constant RC.

For 100% modulation,  $m_a = 1$

$$\therefore \frac{1}{RC} \geq \frac{m_a}{\sqrt{1-m_a^2}} \text{ approaches infinity}$$

It means for  $m_a = 1$ , time constant  $RC = 0$ , i.e.  $C = 0$ . At this condition of no capacitor in the circuit the output contains carrier frequency component of voltage as mentioned at the beginning.

When the percentage modulation is not very large, as is usually the case, the value of  $\sqrt{1 - m_a^2}$  tends to be nearly unity, giving  $\frac{1}{RC} \geq \omega_m \cdot m_a$ .

### 11.5 Detection of Frequency Modulated Wave

The FM detector sometimes called the discriminator performs the function of extracting the modulating voltage from the frequency modulated voltage. In the detection or demodulation process, the discriminator is made to change the frequency deviation of the modulated voltage into an amplitude variation corresponding to the modulating voltage. This conversion from frequency variation to amplitude variation should be linear and efficient. The ideal FM wave is of constant amplitude and hence any variation of amplitude of the received frequency modulated wave is an indication of noise. The detector, therefore, should not respond to any variation of amplitude. Actually the discriminator i.e. the FM detector has to convert the frequency modulated wave of constant amplitude into a voltage possessing both amplitude variation and frequency variation. It means the frequency deviation is used to change the amplitude keeping the inner frequency the same. This latter voltage is then applied to appropriate circuitry which responds to the amplitude variation only but ignores the effect of frequency variation.

#### Slope Detector

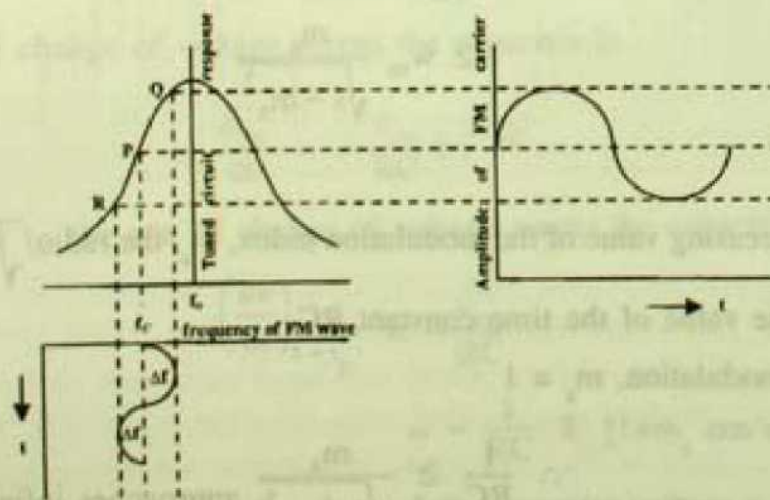
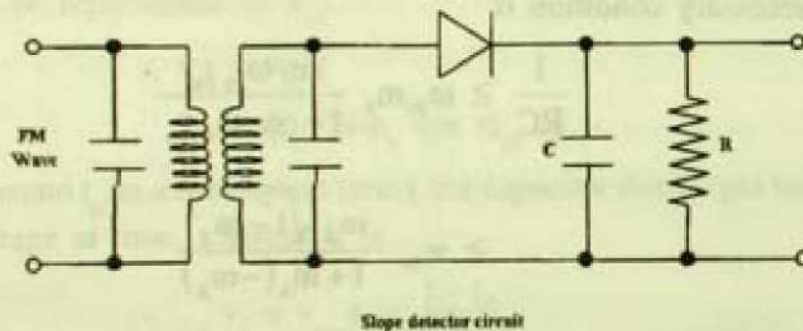


Fig. 11.16 Slope detector circuit and the waveform



The single tuned discriminator circuit consists of parallel resonant circuit tuned to a frequency slightly different from the carrier or centre frequency of FM wave. It is followed by the diode and the filter section in series. The action of the slope detector in converting the frequency modulated wave into a corresponding amplitude variation of the voltage has been shown graphically (Fig. 11.16). In the circuit arrangement, the resonant frequency  $f_0$  is higher than the centre frequency  $f_c$  of the FM wave. The frequency of the modulated voltage varies from  $(f_c - \Delta f)$  to  $(f_c + \Delta f)$ . With a constant amplitude of the modulated voltage having frequency variation from  $(f_c - \Delta f)$  to  $(f_c + \Delta f)$  the output of the tuned circuit will change from the point R to Q respectively. The frequency deviation is thus converted into amplitude variation of the voltage keeping the frequency intact. The amplitude modulated voltage is fed to the diode detector the output of which gives back the modulating voltage wave. If the response curve of the tuned circuit on both sides of the point P lies on a straight line without any curvature, the frequency deviation will produce a linear amplitude variation and it will give a faithful reproduction of modulating voltage. But actually the response curve is never a perfect straight line for a wide range of frequency deviation thus producing distorted output. Moreover, the tuned circuit cannot reject the input amplitude variation, if any of the FM wave giving a proportional output voltage. As a matter of fact the single tuned discriminator is not used in practice due to these basic limitations. Slope detector may be operated on the other side i.e. falling side of the response curve of the tuned circuit. It will give an inverted modulating voltage at the output i.e. the output voltage will decrease with increasing frequency deviation of the FM wave and vice-versa.

It is a difficulty to adjust the circuit properly, since the primary and secondary tuned circuits are to be tuned at a slightly different frequencies.

### Balanced Slope Detector

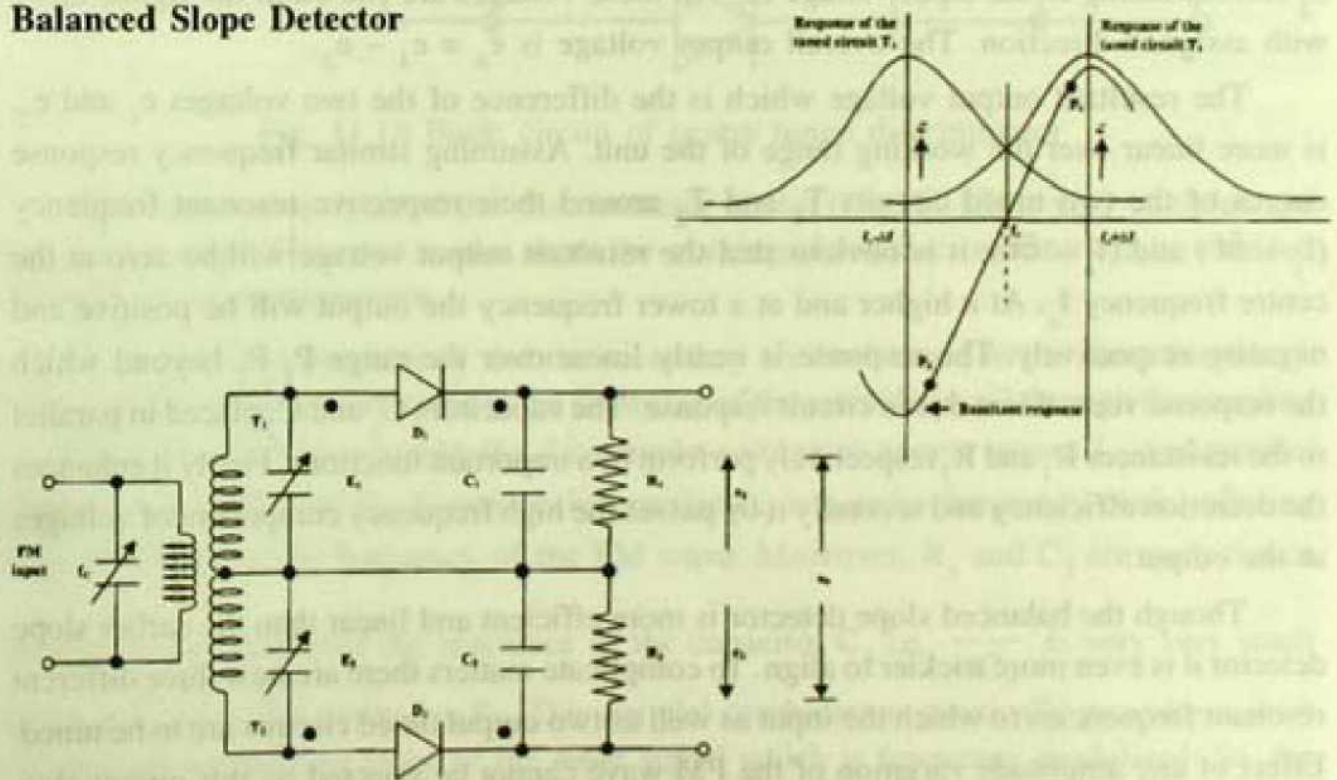


Fig. 11.17 Balanced slope detector circuit and the waveform



A balanced slope detector (Fig. 11.17) uses two tuned circuits  $T_1$  and  $T_2$  in association with their respective diode detectors and filtering arrangements. The primary tuned circuit is tuned at the centre carrier frequency  $f_c$ . Tuned circuit  $T_1$  is tuned at a higher frequency than  $f_c$ , say at  $f_c + \delta f$  where  $\delta f$  is more than the frequency deviation allowed in the frequency modulation process. Similarly the tuned circuit  $T_2$  is tuned at a slightly lower frequency ( $f_c - \delta f$ ) which is lower than centre frequency of the frequency modulated wave. Thus  $f_c$ , the centre frequency of the modulated carrier voltage is at the middle and the two tuned frequencies  $f_c + \delta f$  and  $f_c - \delta f$  are symmetrically placed on the two sides of the frequency  $f_c$ . The input is an FM wave whose frequency is varying with time so the response of the two tank circuits  $T_1$  and  $T_2$  will be different at all frequencies except at  $f_c$  (the centre frequency of the FM wave). As a result when the voltage across the tank circuit  $T_1$  will be of a higher frequency due to input frequency  $f_c + \Delta f$ , the voltage obtained across  $T_2$  will be of a lower value at the frequency  $f_c + \Delta f$  because  $T_2$  is tuned at a much lower frequency  $f_c - \delta f$ . Similarly when the input frequency is going down the centre frequency i.e. at a input frequency say  $f_c - \Delta f$ , the response of the  $T_1$  tuned circuit will be lower while that of  $T_2$  will be higher. Diode  $D_1$  in combination with  $R_1$  and  $C_1$  constitutes one detector arrangement supplied by  $T_1$  while diode  $D_2$  in combination with  $R_2$  and  $C_2$  constitutes the second detector arrangement being fed by  $T_2$ . The first detector is supplied by  $E_1$  and gives a linear output voltage  $e_1$  while the second unit gives an output voltage  $e_2$  corresponding to the input voltage  $E_2$ . All these voltages are shown in the figure 11.17 with assigned direction. The overall output voltage is  $e_o = e_1 - e_2$ .

The resultant output voltage which is the difference of the two voltages  $e_1$  and  $e_2$ , is more linear over the working range of the unit. Assuming similar frequency response curves of the two tuned circuits  $T_1$  and  $T_2$  around their respective resonant frequency ( $f_c + \delta f$ ) and ( $f_c - \delta f$ ), it is obvious that the resultant output voltage will be zero at the centre frequency  $f_c$ . At a higher and at a lower frequency the output will be positive and negative respectively. The response is nearly linear over the range  $P_1$   $P_2$  beyond which the response veers down due to circuit response. The capacitors  $C_1$  and  $C_2$  placed in parallel to the resistances  $R_1$  and  $R_2$  respectively perform two important functions. Firstly it enhances the detection efficiency and secondly it by passes the high frequency component of voltages at the output.

Though the balanced slope detector is more efficient and linear than the earlier slope detector it is even more trickier to align. To complicate matters there are now three different resonant frequencies to which the input as well as two output tuned circuits are to be tuned. Effect of any amplitude variation of the FM wave cannot be rejected by this circuit also. The linearity, although better than that of the single slope detector, is still not perfect.



### Phase Discriminator/ Foster-Seeley Discriminator/ Centre Tuned Discriminator

It is the most commonly used type of discriminator used for detection of frequency modulated wave. It gives the same S shaped response curve at the output although the primary and secondary tuned circuits are all tuned to the identical centre frequency of the incoming frequency modulated wave. Tuning at a single frequency greatly simplifies the alignment. Though very simple in single frequency tuning, this discriminator with its proper and tactful placement of circuit elements provides better linearity than the slope detector discussed so far.

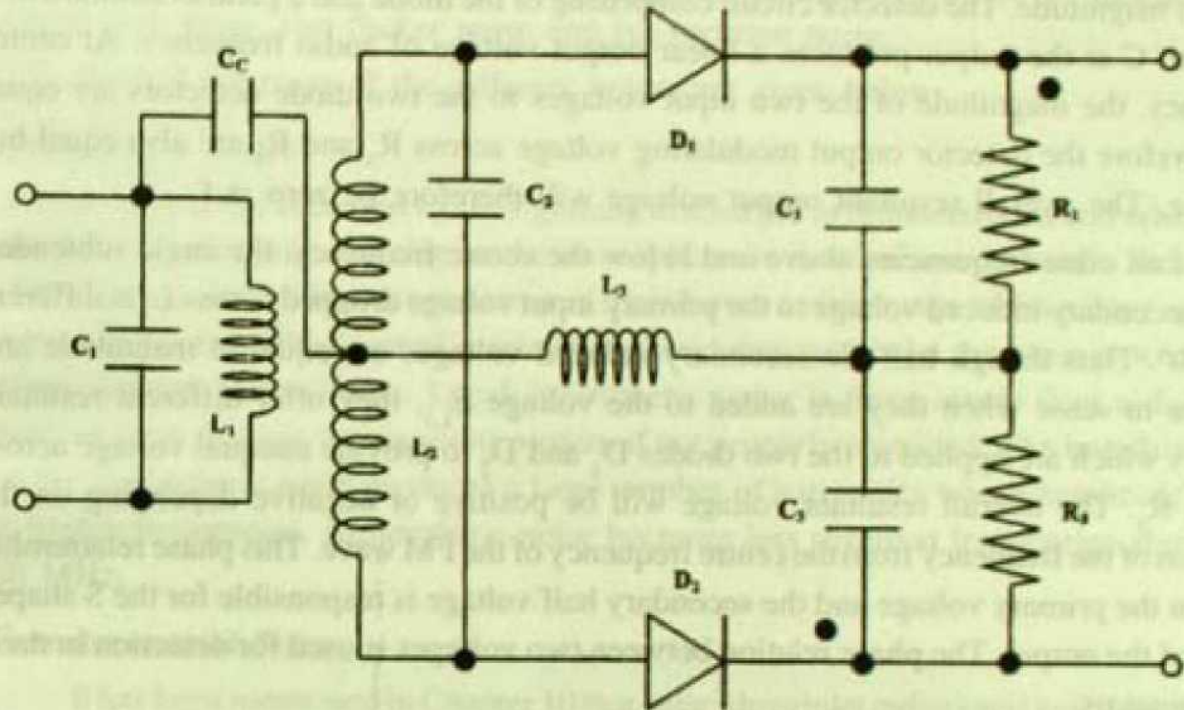


Fig. 11.18 Basic circuit of centre tuned discriminator

The rigorous analysis of this discriminator circuit (Fig. 11.18) is very complex and beyond the scope. However an idea about the working of this discriminator can be obtained through circuit arrangements.

As mentioned the primary and the secondary are both tuned to the same frequency. The primary voltage is connected to the centre tap of the secondary coil through the coupling capacitor  $C_c$  which also blocks the d.c. supply voltage to appear across  $L_3$  and parallel combination of  $R_3$  and  $C_3$ . However, in practice  $L_3$  is a radio frequency choke offering large reactance at the frequency of the FM wave. Moreover,  $R_3$  and  $C_3$  are such that at the frequency concerned the reactance of the capacitor  $C_3$  i.e.  $\frac{1}{\omega C_3}$  is very very small in comparison to the resistance  $R_3$ . This parallel combination practically provides a short circuit path for the frequency of the input signal which is frequency modulated. So, the total input a.c. voltage drops across the choke  $L_3$ .



Secondly, looking to the mutual coupling of the two transformer windings, it is obvious that a voltage will be induced into the secondary. As mentioned earlier the secondary is centre tapped as a result of which half of this induced voltage should be added vectorially with the voltage dropped across  $L_1$ , to get the resultant voltage applied to the input to the diodes  $D_1$  and  $D_2$  respectively. Here is actually the action of the circuit. Just at the centre frequency  $f_c$ , each half of the secondary voltage which is to be added to  $E_{L_1}$  (voltage across  $L_1$ ) with angle  $90^\circ$ , are equal in magnitude but opposite in sense, making the resultant voltages equal in magnitude. The detector circuit comprising of the diode and a parallel combination of  $R$  and  $C$  at the output produces a linear output voltage of audio frequency. At centre frequency, the magnitude of the two input voltages to the two diode detectors are equal and therefore the detector output modulating voltage across  $R_2$  and  $R_1$  are also equal but opposite. The overall resultant output voltage will therefore be zero at  $f_c$ .

At all other frequencies above and below the centre frequency, the angle subtended by the secondary induced voltage to the primary input voltage dropped across  $L_1$  is different from  $90^\circ$ . Thus though half the secondary induced voltages are equal in magnitude and opposite in sense when they are added to the voltage  $E_{L_1}$ , they offer different resultant voltages which are applied to the two diodes  $D_1$  and  $D_2$  to provide unequal voltage across  $R_1$  and  $R_2$ . The overall resultant voltage will be positive or negative depending on the deviation of the frequency from the centre frequency of the FM wave. This phase relationship between the primary voltage and the secondary half voltage is responsible for the S shaped nature of the output. The phase relation between two voltages is used for detection in these discriminators.

## 11.6 Noise

In electronics in the simplest word, the unwanted signal is the noise. Noise may be defined as any unwanted energy which interferes with the perfect reception and reproduction of the wanted signals. It is ever present and limits the performance of virtually every system or device. For example, in radio, noise causes the loud speaker output to be irritating whereas in the case of television reception, it causes the "snow" to be superimposed on the picture. Noise level puts a limit on the lowest signal that can be utilised by an electronic system.

Noise may be classified into different ways. Depending on the circumstances, noise may be classified according to their types, sources, effects on the receiving systems. But it is simpler to categorise them into two groups viz.

- (a) external noise and
- (b) internal noise

The different forms of noise generated outside the receiving system are called the external noise. The important contributors to the external noise are (i) atmospheric noise,



(ii) extra terrestrial noise and (iii) man made noise. The receiving systems with its active and passive elements have no control on the external noise.

On the contrary, the noise generated by the active and passive components of the receiving system is called the internal noise of the system. Such noise is generally random. As this noise is randomly distributed over the entire radio frequency, there is, on the average, as much of it, at any frequency as at any other frequency. Thus the random noise power is proportional to the bandwidth of the receiving system. The important contributors of the internal noise in a receiving system are (i) thermal agitation noise, (ii) shot noise, (iii) transit time noise, (iv) flicker noise and (v) partition noise.

A brief summary of the different noises are given below:

### **Atmospheric Noise**

Atmospheric noise is caused by the lightning discharges in thunderstorms and other natural electrical discharges occurring in the atmosphere. It is impulsive in nature and because of the randomness of its occurrences it is spread over a wide range of frequency spectrum. It propagates from the point of origin to the receiving system by the same process as that of the wanted radio signals. Local atmospheric noise is more severe than a distant one and can even damage the receiving system if not properly protected. The impulsive nature of the atmospheric noise produces a large number of harmonics whose amplitude falls off at higher frequencies. Atmospheric noise becomes less severe at frequencies above about 30 MHz.

### **Extraterrestrial Noise**

It has been mentioned in Chapter 10 that solar ultraviolet radiation is mainly responsible for the formation of ionospheric layers which helps the propagation of short wave radio broadcasting. But occasionally the sun produces much noise jamming radio propagation totally. Solar radio emission usually consists of a basic component and a varying component. The varying component sometimes embodies solar radio bursts which are generated at the active centres on the solar surface. The mechanism of their generation is complex. The bursts of radio waves generated there have a wide range of frequency spectrum and reach the earth in usual time taken by an e.m. wave and produces disturbances in the receiving system. This is solar noise. The solar activity has a 11 year cycle, it has also a long period 100 years cycle. Bursts are known to be generated in and around sunspots, dark patches on the solar surface. Occasionally, charged particles are also emitted from the active region of the sun, they proceed towards polar regions of the earth and produce radio disturbances.

The unwanted radio waves which are received from any star except the sun and also from any other distant radio sources are called cosmic noise. This noise is also called the thermal or black body noise and is distributed fairly uniformly over wide frequency range. Noises are received from our own galaxy (milky way), from other galaxies and also from 'quasars' and 'pulsars'.



## Man-made Noise

Human being produces electrical noise on the earth and this man-made noise is more intense than any other internal or external noise handled by the receiving system. This noise is produced by sources like automobiles, aircraft ignition, electric motors, switching gears, high voltage line leakage and multitudes of other heavy electric machines in the urban, suburban and other industrial areas. This noise is spread approximately over 1 to 600 MHz region.

## Thermal Agitation Noise

The noise generated in a resistance or in the resistive component of an impedance is called the thermal noise or agitation noise or white noise or Johnson noise. The free electrons within a conductor are in random motion when subjected to thermal energy.

The kinetic theory shows that the temperature of a particle is a way of expressing its internal kinetic energy, so the temperature of a body is the measure of statistical rms value of the velocity of the particles in the body. The kinetic energy of these particles approach zero at 0°K. It is thus apparent that the noise power generated in a resistance is proportional to its absolute temperature. Moreover, the noise generated is proportional to the bandwidth over which the noise power is to be considered. Therefore, the noise power

$$P_n \propto TB$$

$$= KTB$$

where,  $P_n$  = maximum noise power output of a resistor

$K$  = Boltzmann's constant =  $1.38 \times 10^{-23}$  J/K

$T$  = absolute temperature

$B$  = bandwidth over which noise power is considered.

It may be thought that there is no voltage across the two terminals of a resistor if the resistor is not connected to any voltage source. This is correct if the measurement is made by a d.c. voltmeter, but it is incorrect if a very sensitive electronic voltmeter is used. The noise generated in the resistor may even produce a large voltage across it, but since it is random, and therefore has a definite rms value but no d.c. component, only an ac meter will register a reading. This noise voltage is caused by the random movement of the electrons within the resistor which constitutes a current. At any instant of time, unequal number of electrons may reach the two ends of the resistor but considering a long period the numbers may be equal. Over a long period of time the imbalance may be reduced, but the rate of arrival of electrons at either end of the resistor varies randomly, so does the potential difference between the two ends of the resistor. Thus a random voltage across the resistor definitely exists. The noise voltage generated across a resistor  $R$  is given by



$$E_n = \sqrt{4KTBR}$$

where  $E_n$  = noise voltage

$R$  = resistance across which the noise voltage is generated.

### Shot Noise

Shot noise is produced by random variation in the arrival of electrons or holes at the collector of a transistor or due to electron arrival at the plate of a vacuum tube. It manifests itself as a randomly varying noise current superimposed on direct current. The audio response of this noise appears as if a shower of lead shots were falling on a metal sheet, hence the name shot noise. Many variables are involved in the generation of shot noise in a semiconductor device like transistor or in a vacuum tube. It is customary to use simplified expression for shot noise except for a diode. The exact expression for the rms value of the diode shot current,  $i_n$ , is given by,

$$i_n = \sqrt{2e i_d B}$$

where  $e$  = charge of an electron =  $1.6 \times 10^{-19}$  C

$i_d$  = direct diode current

and  $B$  = band width.

### Transit Time Noise or High Frequency Noise

In semiconductor devices, if the transit time of the carriers crossing a junction is comparable with the periodic time of the signal, some of the carriers may diffuse back to the source or emitter. It can be shown that this gives rise to an input admittance. Moreover, the conductance component of the admittance increases with frequency. Thus this high frequency noise is frequency dependent. Similar effect occurs in vacuum tubes when the transit time of the electrons in moving from the cathode to the control grid is comparable to the periodic time of the signal.

### Flicker Noise or Low Frequency Noise

At low audio frequencies, below 1 kHz, a poorly understood form of noise called flicker noise is found in transistors. It is also called modulation noise. This noise is proportional to emitter current and junction temperature. This noise is found to be inversely proportional to the frequency and hence its contribution is very small at frequencies above 500 Hz.

### Partition Noise

Partition noise occurs whenever current divides between two or more paths. This noise results from the random fluctuations in the division. It would be expected, therefore, that a diode would be less noisy than a transistor. This partition noise is also found in multigrid tubes. In radio receivers, mixers are more noisy particularly due to high partition noise in the active device irrespective of transistor or vacuum tube.

## 11.7 Signal-to-Noise Ratio

Each and every electronic system will produce its own internal noise due to the various components of the system. At the input of the system the desired signal invariably combines with the external noise. At the successive stages of the system, the total noise at these stages are the combination of external noise and internal noise generated so far. If some where in the flow path, the noise exceeds the signal strength, the wanted signal is masked by the unwanted noise and every information is lost from that stage without further addition in the successive stages of the electronic system. The ratio of signal power to noise power should never be less than unity to avoid masking by the noise. This indicates the importance of signal to noise ratio in a system.

The ratio of the signal power to noise power at the same point of a system may be expressed as,

$$\frac{S}{N} = \frac{P_s}{P_n} = \frac{\frac{E_s^2}{R}}{\frac{E_n^2}{R}} = \left( \frac{E_s}{E_n} \right)^2$$

where  $S = P_s$  = signal power at the point concerned  
 $N = P_n$  = noise power at the point concerned  
 and  $R$  = resistance

$E_s$  and  $E_n$  are respectively the signal voltage and noise voltage at the same point. The equation is a result of simplification and it applies whenever the resistance across which the noise is developed is the same as the resistance across which signal is developed.

Signal to noise ratio  $\left( \frac{S}{N} \right)$  is a very important quantity and it varies from point to point. So, for an amplifier or a network, the signal to noise ratio at the input will differ from the signal to noise ratio at the output. Their relative ratio is usually represented by  $F$  and is expressed by

$$F = \frac{\left( \frac{S}{N} \right) \text{ power ratio at the input}}{\left( \frac{S}{N} \right) \text{ power ratio at the output}}$$

$$= \frac{P_{si}}{P_{ni}} \times \frac{P_{no}}{P_{so}}$$

where  $F$  = noise factor of the amplifier or network  
 $P_{si}$  = signal power available at the input  
 $P_{ni}$  = noise power available at the input



$P_{no}$  = noise power available at the output

$P_{so}$  = signal power available at the output

The  $\frac{S}{N}$  available at the output will always be less than that at the input, since any amplifier or network will add noise. Therefore the noise factor,  $F$ , is a measure of the amount of noise added and  $F$  will always be greater than unity.

The noise factor  $F$  is frequency dependent in many cases and where it is determined at one frequency it is called the "spot noise factor". An average value of  $F$  can also be found, over the frequency range of interest and it is called the "average noise factor",  $F_{av}$ .

### 11.8 Solved Problems

1. The frequency of a carrier voltage is 600 kHz. What will be the side frequencies and bandwidth of the transmitted signal when the carrier voltage is amplitude modulated by an audio modulating voltage of frequency 3 kHz.

Given carrier frequency  $f_c = 600$  kHz

Given modulating frequency  $f_m = 3$  kHz

So, the upper side frequency  $= f_c + f_m = 600$  kHz + 3 kHz = 603 kHz

and the lower side frequency  $= f_c - f_m = 600$  kHz - 3 kHz = 597 kHz

Required bandwidth  $= 2 f_m = 2 \times 3$  kHz = 6 kHz.

2. The peak to peak voltage of an amplitude modulated wave is 4 volts and the minimum dip to dip voltage is 1 volt only. Calculate the percentage modulation of the modulated wave and the amplitude of the unmodulated voltage.

In case of AM, the modulation index

$$m_a = \frac{E_{max} - E_{min}}{E_{max} + E_{min}}$$

$$= \frac{4 - 1}{\frac{4 + 1}{2}} = .6$$

$$\begin{aligned} \text{Percentage modulation} &= 100 m_a \% \\ &= 100 \times .6 \% = 60 \% \end{aligned}$$

$$\text{Moreover, } m_a = \frac{E_{max} - E_{min}}{2E_c}$$

where  $E_c$  is the amplitude of the carrier voltage

$$\therefore E_c = \frac{E_{\max} - E_{\min}}{2m_a} = \frac{(4-1)/2}{2 \times 6} = 1.25 \text{ volt}$$

3. A high frequency carrier voltage  $e_c = 15 \cos 2\pi 10^7 t$  is amplitude modulated by a modulating voltage  $e_m = 5 \cos 2\pi 6 \times 10^3 t$ . Calculate the modulation index produced as well as the frequencies and amplitudes of the side frequencies.

$$\begin{aligned} \text{Modulation index } m_a &= \frac{K_a E_m}{E_c} \\ &= K_a \cdot \frac{5}{15} \end{aligned}$$

Amplitude of the carrier voltage,  $E_c = 15 \text{ V}$  (given)

Amplitude of the modulating voltage,  $E_m = 5 \text{ V}$  (given)

And  $K_a = 1$

$$\therefore m_a = \frac{5}{15} = .33$$

$$\% m_a = 33\%$$

Carrier frequency  $f_c = 10^7 \text{ Hz}$

Modulating frequency  $f_m = 6 \times 10^3 \text{ Hz}$

$$\begin{aligned} \therefore \text{Upper side frequency} &= f_c + f_m \\ &= 10^7 \text{ Hz} + 6 \times 10^3 \text{ Hz} \\ &= 10.006 \text{ MHz} \end{aligned}$$

$$\begin{aligned} \text{Lower side frequency} &= f_c - f_m \\ &= 10^7 \text{ Hz} - 6 \times 10^3 \text{ Hz} \\ &= 9.994 \text{ MHz} \end{aligned}$$

Amplitude of the upper side frequency and lower side frequency

$$\begin{aligned} &= \frac{1}{2} m_a \cdot E_c \\ &= \frac{1}{2} \times \frac{1}{3} \times 15 \\ &= 2.5 \text{ V.} \end{aligned}$$

4. A carrier voltage has the amplitude 200 V and frequency 1 MHz. When it is 40% modulated, calculate the power delivered by this amplitude modulated wave to a load impedance of value  $50 \Omega$ .



The unmodulated carrier power is given by

$$P_c = \frac{E_c^2}{2R}$$

$$= \frac{(200)^2}{2 \times 50} = 400 \text{ W}$$

The total power of the amplitude modulated wave i.e. the power in the carrier and two side bands is given by

$$P_t = P_c + P_{s1} + P_{s2}$$

$$= P_c \left( 1 + \frac{m_a^2}{2} \right)$$

$$= 400 \left( 1 + \frac{(4)^2}{2} \right)$$

$$= 432 \text{ W}$$

5. The frequency swing of an FM wave is 80 kHz. Calculate the frequency modulation index,  $m_f$  for an audio modulating frequency 4 kHz.

Frequency swing in FM = 2 (frequency deviation)

$$\therefore \text{Frequency deviation, } \Delta f = \frac{\text{frequency swing}}{2} = \frac{80 \text{ kHz}}{2}$$

$$= 40 \text{ kHz.}$$

$$\text{Frequency modulation index } m_f = \frac{\text{frequency deviation}}{\text{modulating frequency}}$$

$$= \frac{40 \text{ kHz}}{4 \text{ kHz}}$$

$$= 10.$$

6. The 100 MHz centre frequency of a carrier voltage is frequency modulated by an audio frequency 15 KHz producing a maximum frequency deviation 60 KHz. Calculate the value of frequency modulation index. Also calculate five significant pairs of side frequencies and the channel width required for their transmission.

The frequency modulation index  $m_f$  is given by,

$$m_f = \frac{\Delta f}{f_m} = \frac{60 \text{ kHz}}{15 \text{ kHz}} = 4$$

Five significant side frequency pairs are

$$100 \text{ MHz} \pm 15 \text{ kHz,} = 100.015 \text{ MHz, } 99.985 \text{ MHz}$$

$$100 \text{ MHz} \pm 30 \text{ kHz,} = 100.03 \text{ MHz, } 99.97 \text{ MHz}$$

$$100 \text{ MHz} \pm 45 \text{ kHz}, = 100.045 \text{ MHz}, 99.955 \text{ MHz}$$

$$100 \text{ MHz} \pm 60 \text{ kHz}, = 100.06 \text{ MHz}, 99.94 \text{ MHz}$$

and  $100 \text{ MHz} \pm 75 \text{ kHz}, = 100.075 \text{ MHz}, 99.925 \text{ MHz}$

The channel width required for transmission is

$$2 \times 5 \times 15 \text{ kHz} = 150 \text{ kHz}$$

7. The RC parallel combination of a linear diode detector uses  $.22 \text{ M}\Omega$  and  $250 \text{ pf}$  respectively. If 30% AM wave is fed to the detector, what is the highest audio modulation frequency which can be detected without distortion ?

The maximum audio frequency will be

$$\begin{aligned} f_m &= \frac{\omega_m}{2\pi} = \frac{1}{2\pi} \cdot \frac{1}{m_a RC} \\ &= \frac{1}{2\pi \times .3 \times .22 \times 10^6 \times 250 \times 10^{-12}} \text{ Hz} \\ &= 9.64 \text{ Hz} \end{aligned}$$

## 11.9 Questions and Problems

### Questions

1. What is modulation ? What is the need for modulation ?
2. What are important types of modulation ? What is the basic difference among them ?
3. Define amplitude modulation. What is its depth of modulation ?
4. Give the mathematical expression for amplitude modulation.
5. Explain frequency spectrum of AM wave.
6. Draw the waveform of an AM wave. Show how to obtain the modulation index from the waveform.
7. Find an expression for sideband powers in term of total power and modulation index.
8. Show that the total power for 100% modulated AM wave is 1.5 times the unmodulated carrier power.
9. Draw the circuit and explain how amplitude modulated wave can be produced.
10. What are the limitations of amplitude modulation ?
11. What is frequency modulation ? Derive an expression for an FM wave with sinusoidal voltages.
12. Draw the waveforms of an FM wave produced with sinusoidal voltages.
13. Write a note on the frequency spectrum of an FM wave.
14. Comment on the sidebands present in an FM wave. Compare these sidebands with those of AM wave.
15. Compare frequency modulation with amplitude modulation.



16. What is demodulation or detection ? Explain it.
17. Draw the circuit of an envelope detector and explain its operation.
18. Find the value of the highest modulating frequency which can be detected by linear diode detector without clipping.
19. Discuss the types of noise which hinders reception.
20. Write a note on signal to noise ratio.

### Problems

1. A carrier voltage of 10 V amplitude and frequency 1000 kHz is amplitude modulated by a sine wave of 4 V amplitude and frequency 10 kHz. Determine the modulation index and sketch the waveform.
2. A 5 kHz audio signal of amplitude 20 V, amplitude modulates a carrier voltage of 50 V and frequency 10 MHz. Determine the value of the side frequencies and their amplitudes.
3. A carrier power of 2 kW is amplitude modulated by an audio signal. If the percentage modulation is 40%, determine the total power radiated. What is the power in each side frequency ?
4. In a 50% amplitude modulated wave what will be the power saving if the carrier and one of the sidebands are suppressed ?
5. A broadcast AM transmitter radiates 50 kW carrier power. What will be the radiated power at 50% modulation ?
6. What is the frequency modulation index of an FM wave which has a frequency swing of 150 kHz when the modulation frequency is 10 kHz ?
7. In an FM system, if modulation index  $m_f$  is doubled by halving the modulating frequency, what will be effect on the maximum frequency deviation ?
8. A 6 kHz audio signal frequency modulates a carrier of 90 MHz. If the frequency deviation is 21 kHz, what is the frequency modulation index ?
9. A transmitter radiates 40 kW with an unmodulated carrier wave and 52 kW with modulated wave. What type of modulation it is ? What is the value of modulation depth ?
10. A 50% amplitude modulated wave is fed to a diode detector. The frequency of the modulating voltage varies from 300 Hz to 5 kHz. If the load resistance of the diode detector is  $.25 \text{ M}\Omega$ , find the value of the shunt capacitor to be used for distortion free detection.

### 11.10 References

1. Applied Electronics, Truman S. Gray, Asia Publishing House
2. Applied Electronics, G.K. Mithal, Khanna Publishers, Delhi
3. Electronic Communication System, George Kennedy, Mc GrawHill, Kogakusha Ltd.
4. Electronic Communication, D.Roddy and J.Coolen, Prentice Hall of India Pvt. Ltd.
5. Electronic Fundamentals and Applications, D.Chattopadhyay and P.C.Raksht, New Age International Publishers.



## **Chapter 12**

### **Signal Transmission Through Media**

#### **12.1 Introduction**

The purpose of having a transmission medium is to transport information from one point to another. The information may be voice and video signals or bit streams of data. Various physical media are used for transmission and each one has its own place in terms of bandwidth, propagation delay, cost and ease of installation.

Media are grouped into two broad classes (a) guided media and (b) unguided media. The class of guided media consists of copper wire and optical fibre of various types. The unguided media are links through the atmosphere which carry radio waves and line of sight laser beam. The guided media is extensively used in telephone communication and in local area computer network whereas the unguided media are used in inter-county links including space based satellite communications.

Point to point link, as the name implies, is the system for transmission of digital or analog signals from one point to another using guided or unguided media. The link length may vary from less than a kilometer termed as short haul to several hundreds of kilometers termed as long haul system. If the link length is too long a repeater can be used to regenerate the loss and distortion of signal due to attenuation and dispersion. Broadcast and distributed network is required for transmission of same information to many users. Examples include local-loop distribution and broadcasting of video channels over cable television to many users. For integrated-service-digital-network (ISDN), two network topologies are generally used. They are called hub and bus topology. In hub topology, channel distribution takes place at central locations (or hub) where an automated cross-connection facility switches the channels. In case of bus topology, a single medium carries the multi channel signals throughout the area of service.

#### **12.2 Twisted pair and coaxial copper cable guided media**

The oldest and most common transmission medium is twisted pair of insulated copper wires, typically about 1 mm in diameter. It may be remembered that two parallel wires constitute a simple antenna and therefore may act as a source of electromagnetic interference. The wires are twisted in a helical form to reduce electrical interference from similar pair in close vicinity. The most common application of twisted pair is in the telephone and in the computer communication systems. The transmission bandwidth of the medium depends



on the thickness of the copper wire and on the distance between the source and the receiver of information. Twisted pair used for computer communication is generally known as UTP (Unshielded Twisted Pair). UTP consists of four pairs of insulated twisted pairs in one sheath.

A coaxial cable consists of a stiff copper wire as core surrounded by an insulating material similar to the twisted pair. However, in coaxial copper cable, a cylindrical conductor of woven braided mesh encases the insulator. The outer conducting mesh is covered in a protective plastic sheath and is often grounded. The view is shown in figure 12.1. Though the bandwidth of the signal to be transmitted depends on the length, yet a 1 km length of cable can sustain a data rate of 1 Giga bits per second.

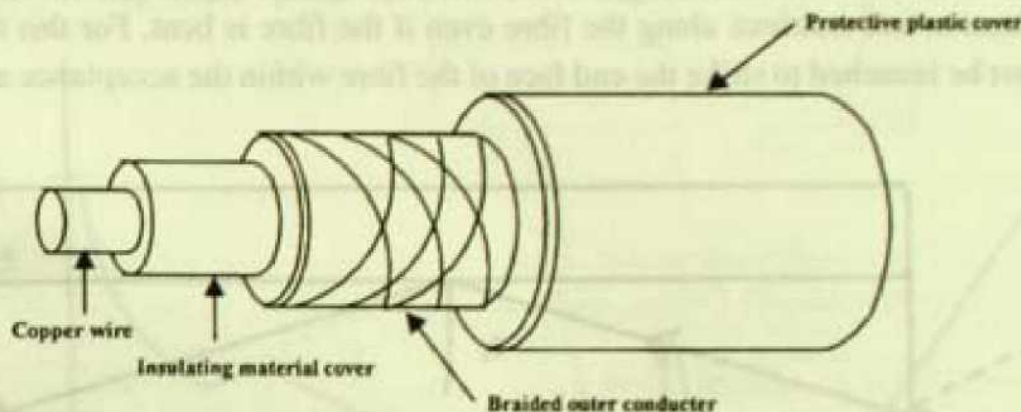


Fig. 12.1 A Typical coaxial cable

Coaxial copper cables are of two categories: base band and broad band coaxial cables. Since early dipole antennas had an impedance of 300 ohms and an impedance matching of 4:1 was used, the base band coaxial copper cable is known as 50 ohm cable. The 50 ohm cable is generally used for digital data transmission. The other kind of coaxial cable, known as broadband cable or 75 ohm cable is used for analog and video transmission. In telephone terminology any frequency higher than 4 kHz is usually termed as broad band signal and in computer communication system broadband cable means any cable network which is used for analog transmission. The broadband cable is therefore used to transmit TV signals up to 450 MHz over a length of 10 to 100 km. Technically the broadband cables are inferior to the base band cables and they often come in a configuration of a pair of cables in a single sheath and are made with thinner copper wires.

### 12.3 Optical fibre

The optical fibres are used as transmission media and have the advantages of (a) high bandwidth ranging from 10 MHz-km to over 1 THz-km, (b) low attenuation around 0.1 dB/km to 3 dB/km, (c) electrical immunity i.e. no radio frequency or electromagnetic interference, (d) security— cannot be easily tapped and has no cross talk (e) flexibility— can be bent to radii of few cms., (f) light weight— under 3g per m and (g) sustenance of higher bit rate ranging from 90 Mb/s to 2.5 Gb/s.



### 12.3.1 Geometric optics of glass fibres

When light propagates in a glass of refractive index  $n$ , it moves more slowly by a factor of  $n$  than in free space. Considering that the velocity of light in free space is about 300,00 km/s, the velocity of light in glass of refractive index 1.5 is 200,000 km/s or 200m/ $\mu$ s. If light emerging from a glass with higher refractive index  $n_2$  is incident on a glass of lower refractive index  $n_1$ , it is totally reflected, provided it strikes the interface below a critical angle  $\theta_c$  (figure 12.2). In case of simplest optical fibre, the *step index fibre*, a circular glass core 10 to 50 micron in diameter is surrounded by a glass cladding whose refractive index is approximately 1% lower than the core refractive index. All light coupled into the fibre by no more than the *critical angle* is therefore guided along the fibre due to total internal reflection and continue along the fibre even if the fibre is bent. For this to occur, the light must be launched to strike the end face of the fibre within the acceptance angle  $\theta_a$ .

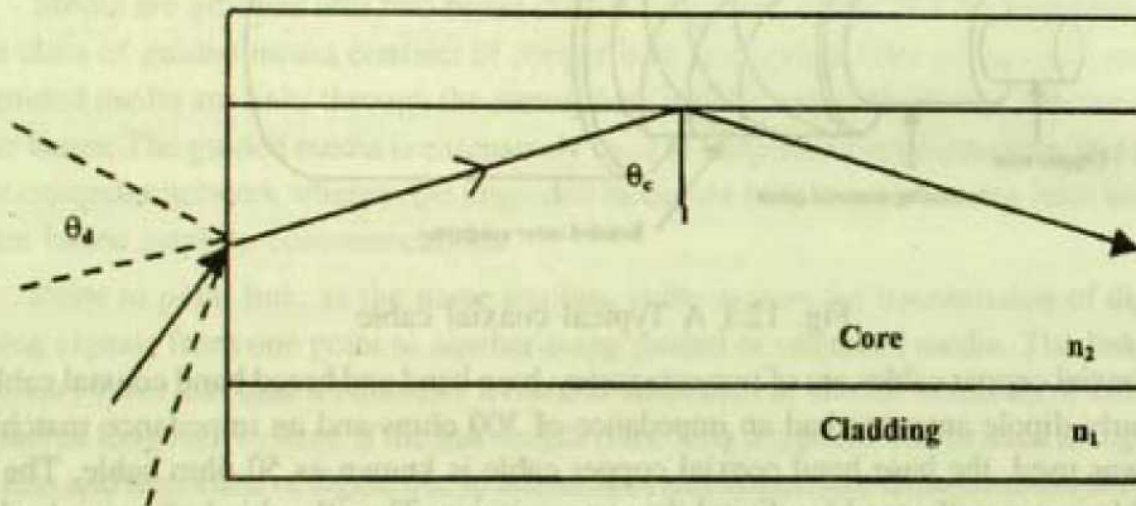


Fig. 12.2 Cutaway view of optical fibre

The sine of the acceptance angle is known as the numerical aperture of optical fibre  $N_A$ . By definition,

$$N_A = [n_1^2 - n_2^2]^{1/2} = n_1 [2\Delta]^{1/2} \quad (12.1)$$

Where  $\Delta$  is the relative refractive index difference and is given by

$$\Delta = [n_1^2 - n_2^2] / 2n_1^2 \quad (12.2)$$

The optical power launched in the fibre decays along the length of the fibre according to an exponential law. The property of the fibre known as the *attenuation coefficient* indicates the loss of optical power in the fibre during propagation. Attenuation coefficient is also a function of the wavelength of light being propagated. There are three principal attenuation mechanisms in fibre. They are absorption, scattering and radiative loss. Radiative loss is



generally kept small by using a sufficiently thick cladding. In the absence of any such loss, the fundamental attenuation mechanism is Rayleigh scattering from irregular glass structure which results in refractive index fluctuation over a small distance compared to wavelength.

This leads to a scattering loss ( $=0.9/\lambda^4$  dB/km) for a reasonably good fibre. Lower attenuation values are achieved around wavelengths of 1.3 and 1.55 micron which are therefore the most suitable wavelengths for long distance transmission. The graph of figure 12.3 shows the fibre loss (attenuation) at various wavelengths. It also shows that the fibre optic medium should be operated either at 1.3 or at 1.55 micron window of wavelength.

One of the major disadvantages of step index optical fibre is the broadening of optical pulse known as *dispersion* (pulse spreading). Reason for this is that the light propagates in the core of step index optical fibre at various angles relative to the fibre axis resulting

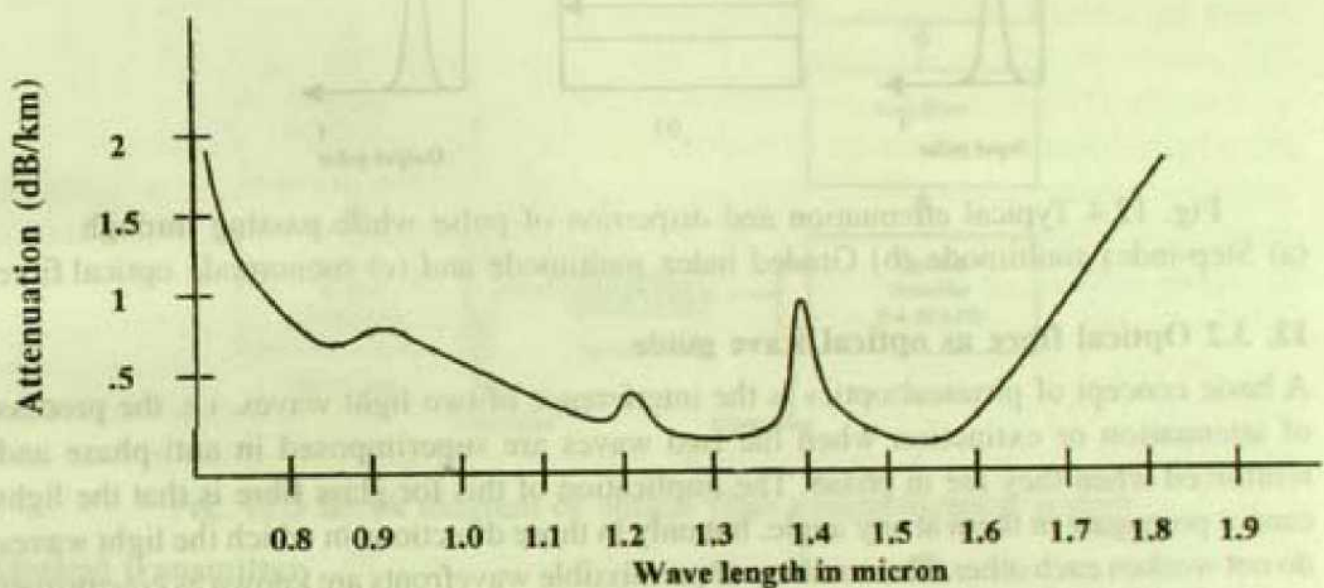


Fig. 12.3 Attenuation in a fibre at various wavelength

in different path lengths. As a result the width of initial short pulse increases steadily as it traverses in the fibre and the pulse appearing at the end of the fibre becomes increasingly spread out. This aspect of fibre behavior is compared to a low pass filter and thus restricts the transmission of wide band signal. This disadvantage of pulse spreading is largely avoided if the refractive index in the fibre core is made to decrease parabolically with radial distance from a maximum at axis to a minimum at the core boundary. In such a *graded index fibre*, the light follows an undulating course instead of zigzag path. The refractive index profile can also be tailor-made so as to give almost zero dispersion of input pulse. For step index fibre, the lowest dispersion per unit length is approximately given by  $\delta t/L = n_1 \Delta/c$  and for graded index fibre the dispersion is  $\delta t/L = n_1 \Delta^2/10c$  where  $c$  is the velocity of light. Therefore for  $\Delta=0.01$ , the dispersion in graded index fibre is 1000 times shorter than that in step index fibre. Figure 12.4 shows the paths in different optical fibres and the effects of attenuation and dispersion on the light path.



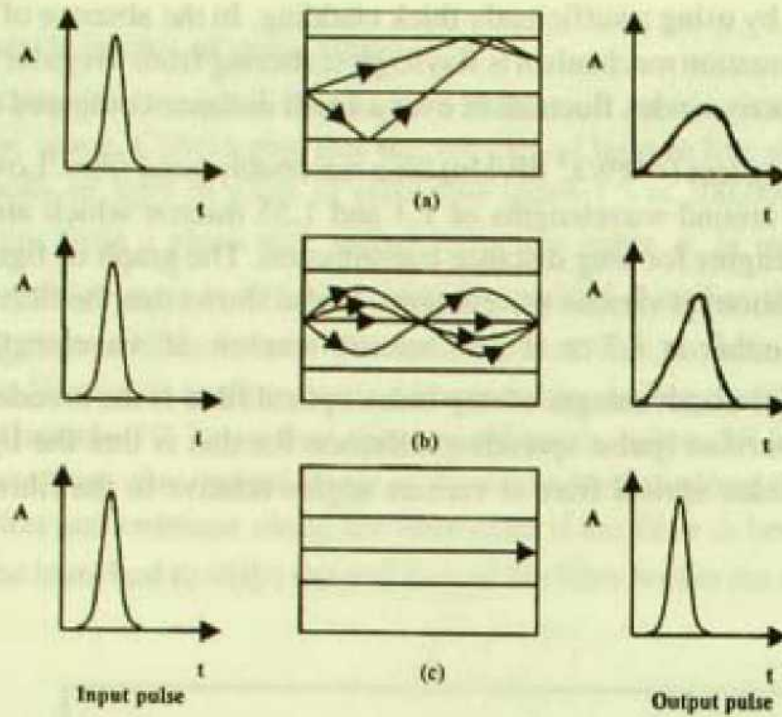


Fig. 12.4 Typical attenuation and dispersion of pulse while passing through (a) Step-index multimode (b) Graded index multimode and (c) monomode optical fibre

### 12. 3.2 Optical fibre as optical wave guide

A basic concept of physical optics is the interference of two light waves, i.e. the process of attenuation or extinction when the two waves are superimposed in anti-phase and reinforced when they are in phase. The implication of this for glass fibre is that the light cannot propagate in them at any angle, but only in those directions in which the light waves do not weaken each other. The numbers of permissible wavefronts are known as *eigenwaves* or *modes* of the fibre. The numbers of modes are very large in multimode fibre because the core diameter is very large in comparison to the wavelength of the light. Therefore the delay difference between light components with different propagation path controlling the transmission bandwidth of the fibre can be expressed as delay difference between different modes. A parameter *v-number* of optical fibre sometimes is used to indicate the number of propagating modes, where

$$v = 2\Delta\pi a/\lambda \quad (12.3)$$

By reducing the core diameter of the optical fibre, the number of propagating modes can be decreased until finally a single mode is left at  $v < 2.4$ . A fibre of this type is known as *single mode fibre* in contrast to the *multimode fibre* having large  $v$  number.

The main advantage of single mode fibre is its large transmission bandwidth, because no delay difference can occur. However, the bandwidth of single mode fibre is essentially limited by an effect known as *material dispersion*. Since the refractive index of glass is wavelength dependent and the light from the source has certain spectral width, a delay difference (intra-modal dispersion) also occurs even in a single mode fibre. Material dispersion is also an important factor in graded index fibre.



### 12.3.3 Communication system using optical fibre as media

A basic schematic diagram of an optical fibre communication system is shown in figure 12.5. There are three basic parts in a typical optical fibre communication system. They are (a) the optical fibre as transmission medium, (b) the optical transmitter and (c) the optical receiver. Some passive optical components and interconnection elements such as connectors and couplers are also used. Electronic systems are used in repeater, multiplexer, coder, decoder and supervisory circuits.

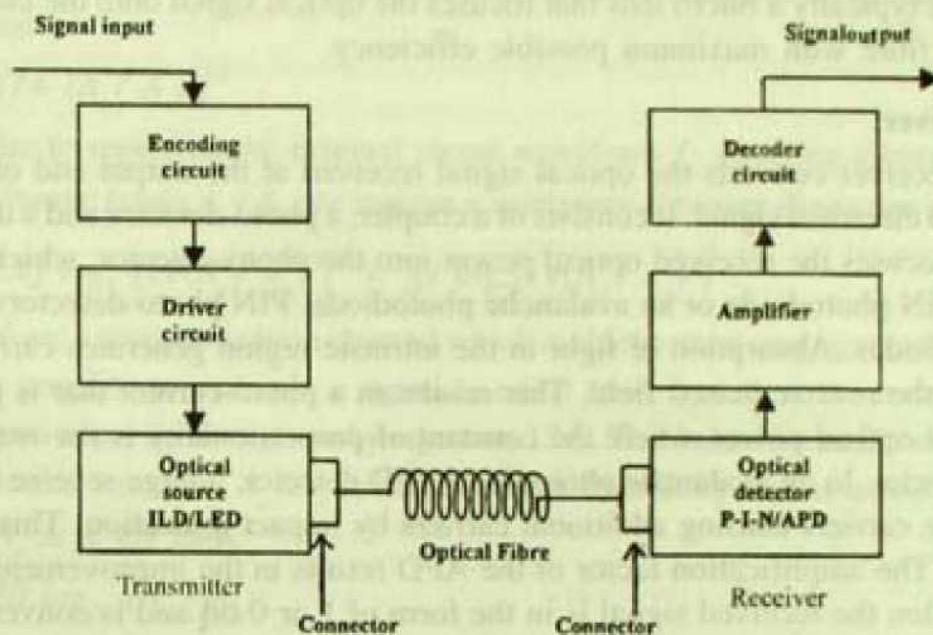


Fig. 12.5 Block diagram of optical fibre communication system

#### Optical transmitter

The role of optical transmitter is to convert the electrical signal into the optical signal and to launch the optical signal into an optical fibre. It consists of an optical source, a modulator and an optical channel coupler. Semiconductor laser (injection laser diode) or light emitting diode is used as optical source. In laser, the population inversion between the ground and excited states results in stimulated emission. In semiconductor laser diode (ILD) this radiation is guided within the active region of the laser and is reflected at the face ends. The combination of feedback and gain results in oscillation when the gain exceeds a threshold value. The spectral width of the ILD is lesser than the LED. The fine line width of a laser results from the change in phase of the optical field. If the intensity of the laser is changed this phase fluctuation gives rise to a change in the frequency of the laser (chirp). The modulation bandwidth of laser which may go to 10 GHz, is determined by resonance frequency caused by the interaction of photon and electron concentrations.

LEDs are forward biased PN junctions in which carrier recombination results in spontaneous emission at a wavelength corresponding to the band gap energy of the semiconductor. Although several milliwatt may be radiated from the LEDs, the radiation is over a wide angular range and consequently there is a large coupling loss from an LED to a fibre. In



addition to radiating over a large angle, the LED radiation has a large spectral width. Owing to the long time constant for the spontaneous emission, the modulation band width of LEDs are generally limited to several nanoseconds.

The optical signal is generated by modulating the optical carrier wave. Injection laser diodes or light emitting diodes can be modulated directly by varying injection current. Since light emitting diodes are limited in their modulation capability, the laser diodes are preferred as source. The bit rate however is rather limited by the electronics than by the light sources. The coupler is typically a micro lens that focuses the optical signal onto the entrance plane of an optical fibre with maximum possible efficiency.

### Optical receiver

The optical receiver converts the optical signal received at the output end of the optical fiber back into electrical signal. It consists of a coupler, a photo-detector and a demodulator. The coupler focuses the received optical power into the photo-detector, which may either be a simple PIN photodiode or an avalanche photodiode. PIN photo-detectors are reverse biased photodiodes. Absorption of light in the intrinsic region generates carriers that are swept out by the reverse biased field. This results in a photo-current that is proportional to the incident optical power, where the constant of proportionality is the *responsivity* of the photo-detector. In an avalanche photodiode APD detector, a large reverse bias voltage accelerates the carriers causing additional carriers by impact ionization. This results in a large current. The amplification factor of the APD results in the improvement of receiver sensitivity. Often the received signal is in the form of 1 or 0 bit and is converted directly into electrical current. The accuracy of the detection is dependent on the signal to noise ratio of the detector. The detector may be integrated with an electronic amplifier for higher gain.

## 12.4 Channel capacity

The purpose of any communication system is to transmit information over a required distance. Therefore, the performance of the system is assessed in terms of the amount of information that may be carried and the distance over which it is sent without repeater. In order to consider the capacity of a channel it is necessary to know how the information is quantified.

While transmitting an analog signal, the first step is to convert the signal in a sequence of binary levels by sampling the analog signal at regular discrete intervals of time known as sampling period  $T (=1/f)$ . According to the sampling theorem it is necessary that the sampling frequency  $f_s$  should be more than twice the highest frequency contained in the original analog signal  $f_m$ . When this condition is satisfied, the original signal is recovered from the sampled signal by passing the waveform through a low-pass filter. The frequency range 0 to  $f_m$  is the bandwidth,  $\Delta f$ , of the signal and thus the sampling frequency  $f_s > 2\Delta f$ .

Each sampled level is then allocated to one of a finite number of amplitude levels. In practice, there is always a random fluctuation present which is superimposed on the signal. This random fluctuation is known as noise. The ratio of maximum signal amplitude



$A_s$  to the r.m.s. noise amplitude  $A_n$  determines the number of levels needed to give sufficiently precise representation of the original signal. If  $m$  is the number of levels chosen, then each sampled value requires  $N$  binary digits to encode it, where

$$N = \log_2 m \quad (12.4)$$

The binary digital signal  $N$  is then decoded back to produce the original signal  $f_m$ . The error produced by the quantization of the sampled amplitude leads to additional noise, known as quantization noise. This noise is less than or comparable to the original noise provided that

$$m > [1 + (A_s / A_n)^2]^{1/2} \quad (12.5)$$

In order to represent the original signal waveform  $f_m$  covering a bandwidth  $\Delta_f$  and having a dynamic range  $A_s / A_n$  we require a minimum  $B$  binary digits per seconds where,

$$B = 2\Delta f \log_2 [1 + (A_s / A_n)^2]^{1/2} = \Delta f \log_2 [1 + (A_s / A_n)^2] \quad (12.6)$$

Therefore, a communication channel which is able to carry analog signals of bandwidth  $\Delta_f$  and it is possible to maintain a signal to noise ratio  $A_s / A_n$  at the receiver, is said to have a channel capacity  $B$  as given by equation 12.6. This is known as Shannon's criterion. In most practical systems,  $A_s / A_n$  is much greater than unity and is usually expressed in dB. If  $A_s / A_n$  is  $X$  dB then,

$$X = 20 \log_{10} (A_s / A_n) \quad (12.7)$$

$$B = \log_2 10 X \Delta_f \quad (12.8)$$

Where  $B$  is expressed in bit per second,  $X$  in dB and  $\Delta_f$  in Hz.

We may estimate the channel capacity required for transmission of a voice channel via a telephone system. We need to transmit a frequency of 300 Hz to 3.4 kHz, which is the baseband signal of voice. The overall signal to noise ratio as specified by most of the telephone companies is 30 dB (i.e.  $A_s / A_n = 31.6$ ). Direct application of Shannon's criterion gives the required bit rate as 30 kb/s with a minimum sampling frequency of 6 kHz. Each sampled amplitude requires at least 5 bit per sample. In practice, a digital pulse code modulated telephone voice channel operates at 64 kb/s. The analog waveform is sampled at 125  $\mu$ s interval ( $f_s = 8$  kHz) and each sample is encoded into a 8 bit word.

Present day TV transmission utilizes VHF/UHF waveband (30-30000 kHz) while normal radio broadcast utilizes lower frequencies (300 kHz to 30 MHz) of the electromagnetic spectrum. The reason for this is that the information content of the TV signals cannot be sent efficiently without loss of information and distortion using radio frequencies as carrier. Indeed, higher one goes up in the electromagnetic spectrum in frequency scale, higher would be the information carrying capacity of a communication system. Since the optical beams have the frequencies in the range of  $10^{14}$ – $10^{15}$  Hz, the use of such beams as carrier would imply a tremendously large increase in the information capacity of the system. However, a light beam (even a laser beam) carrying signals cannot be sent in open atmosphere through



a long distance because of severe attenuation and dispersion due to absorption and scattering by the atmosphere. Therefore, a medium is better if the optical fibre has been developed for carrying light beam with very small attenuation and dispersion.

It is interesting to compare the performance of a coaxial cable based communication channel with that of an optical fibre based channel. At a bit rate of 100 Mb/s, the high quality coaxial cable based communication system having a transmitter power of 1 W (+30dB) and required receiver power of 30 pW (-75dB) would sustain a repeater less distance of 5 km, assuming a signal attenuation of 100 dB. Then the bit rate X distance product is 0.5 Gb/s-km. In contrast in a optical fibre based communication system operating at the same 100 Mb/s bit rate, the optical power transmitted may be taken as 0.5 mW (-3dBm) and the power level required at the receiver is 30nW (-45dBm). Assuming a fibre attenuation of 0.5 dB/km and power margin of 10 dB the achievable link distance would be 64 km. This would give bit rate X distance product as 6.4 GB/s-km.

Table 12.1 shows the number of telephone and TV channels which can be transmitted using different existing technologies.

**Table 12.1** Channels of telephone and TV

Technology	Frequency band	Number of Telephone channels	Number of TV channels
UHF	300–3,000 MHz	10,000	10
Microwave	$3 \times 10^9$ – $10^{12}$ Hz	100,000	100
Optical	$5 \times 10^{13}$ – $10^{15}$ Hz	$10^8$	$10^5$

## 12.5 Microwave communication

Microwave radio transmission is distinguished from other radio applications by its frequency range and by the use of higher directive antennas. At present, the lowest frequency that is allotted for microwave transmission is 1.71 GHz; the highest is 40 GHz. These frequencies are used for microwave transmission because it is practical to focus the radio energy into a beam and communicate a higher percentage of that energy at a receiving location. For example, a 2 m parabolic antenna focuses a 2 GHz radio wave into a beam width of about  $3^\circ$  compared to a  $30^\circ$  beam width at a frequency of 200 MHz.

There are two basic types of microwave communication systems. They are: line-of-sight (LOS) system, which uses relatively low transmitter power over a link length of 17–85 km for ground based communication system. The other type is the over-the-horizon system that uses high transmitter power up to 50 kW or more for beyond the horizon paths (i.e. from 85–1190 km).

Basic microwave system is shown in figure 12.8. In the figure the major distinguishing blocks perform modulator demodulator functions. The microwave frequency source,



frequency conversion and amplification functions are similar to those used in radio communication.

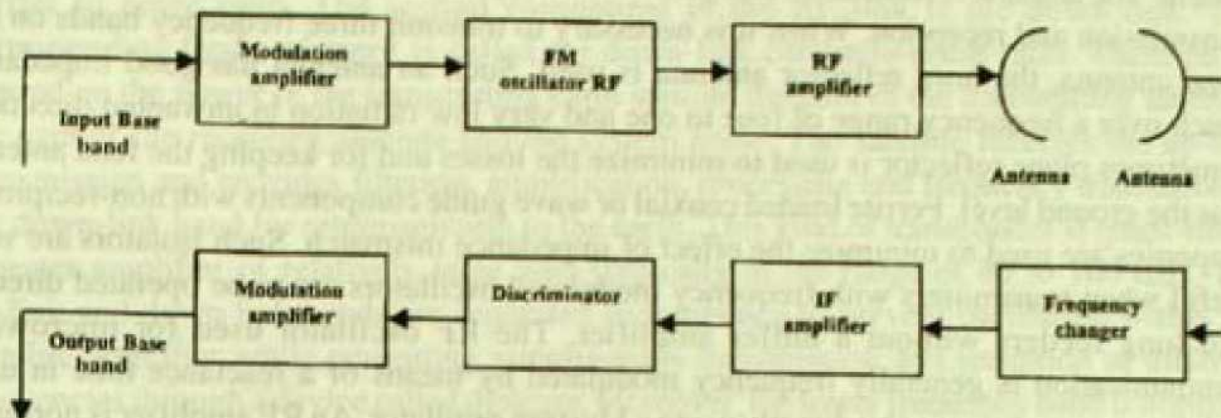


Fig. 12.6 Block diagram of basic microwave communication system

The analog methods of microwave transmission are frequency modulation (FM) and single sideband amplitude modulation. FM used in microwave point to point link has low deviation of the carrier, in contrast to the high-deviation ratio FM used for radio broadcasting. The low deviation causes the higher order modulation sidebands to be substantially below the first order sidebands, so that the FM system is effectively a double side band modulation. The frequency modulated signal has a constant envelop which permits the use of nonlinear class C amplifier. The modulation applied is most commonly used frequency multiplexed set of voice or video channels.

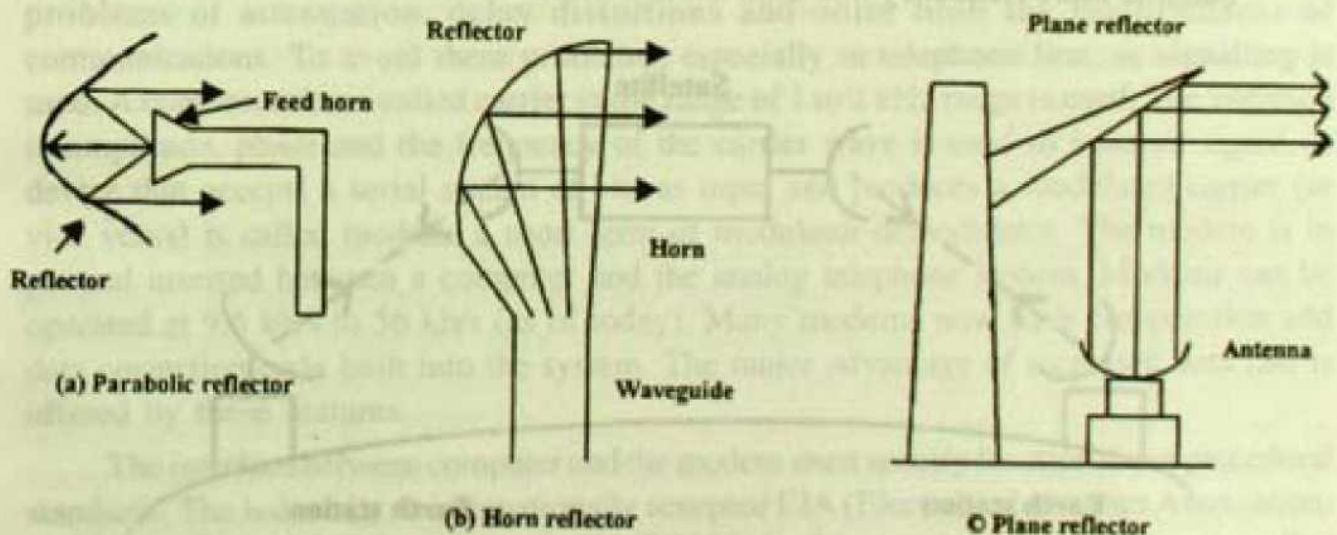


Fig. 12.7 Typical antenna configuration used

The antenna used in microwave communication is shown in figure 12.7. Parabolic reflectors are most commonly used antennas in microwave communication systems. They



are fed by wave guides or horns placed at the focus. The reflectors are about 0.9 to 1.8 meters in diameter for mobile equipment and about 3.6 to 4.5 meter in diameter for fixed system. The antenna is sometimes energized with two planes of polarization for simultaneous transmission and reception. When it is necessary to transmit three frequency bands on the same antenna, the horn reflector antenna is used. Such an antenna has good impedance match over a frequency range of four to one and very low radiation in unwanted direction. Sometimes plane reflector is used to minimize the losses and for keeping the feed antenna near the ground level. Ferrite loaded coaxial or wave guide components with non-reciprocal properties are used to minimize the effect of impedance mismatch. Such isolators are very useful when transmitters with frequency modulated oscillators are to be operated directly into long feeders without a buffer amplifier. The RF oscillator used for microwave communication is generally frequency modulated by means of a reactance tube in ultra high frequency system or by directly using a klystron oscillator. An RF amplifier is normally used between the FM oscillator and the antenna to give higher power output as well as to isolate the oscillator from the feeder reflections. Generally, terminal receivers are of superheterodyne type with a frequency stabilized local oscillator.

### 12.6 Satellite communication

As illustrated in figure 12.8, a communication satellite operates as a distant line-of-sight microwave repeater providing communication services among multiple earth stations in various geographic locations. Most of the communication satellites are placed in geo stationary orbits. Such a satellite remains fixed in apparent position relative to the earth. The satellite's period of revolution is synchronized with that of the earth in inertial space and in close approximation this period is equal to sidereal day.

Satellite radio frequency link establishes a mode of communication between a earth

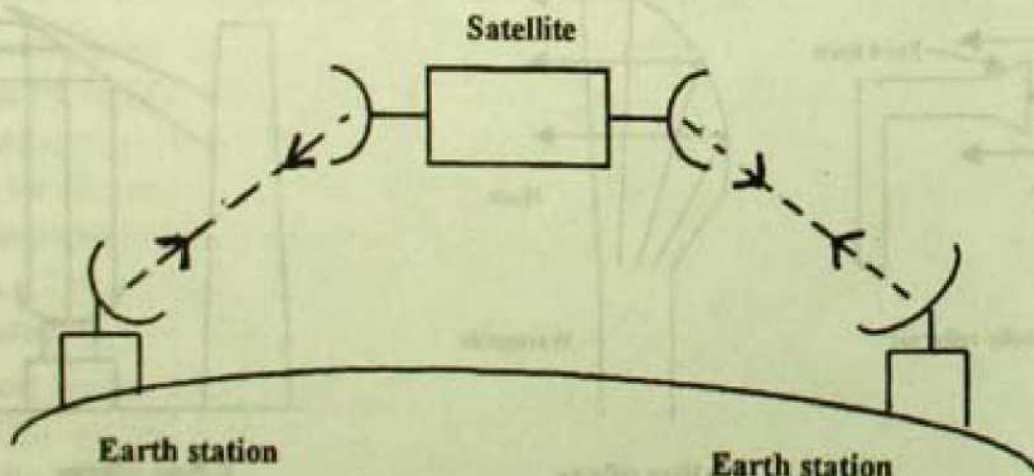


Fig.12.8 A typical satellite communication system

based transmitter and a receiver using satellite as a repeater. The channel carrying capacity of the link is typically expressed in terms of vice channels per transponder and is directly



related to the overall available carrier to noise ratio. Three basic elements are considered in the RF link. The first is the uplink, representing the channel from the transmitting earth station to the satellite. The quality of this link is usually expressed in terms of the uplink carrier to noise ratio. The second component in the RF link is the down link. The corresponding figure of merit is called the down link carrier-to-noise ratio. These ratios depend on the power of the transmitting earth station, the gain of the transmitting antenna, the gain of the receiving antenna and the system noise. The satellite receives the up-link transmission and provides filtering, amplification, processing and frequency translation to the down-link band for retransmission to the earth. This kind of transponder is quasi-linear repeater amplifier of relatively large gain (typically in the range of 80 to 100 dB). The up-link and down-link bands are separated in frequency to prevent oscillation within the satellite amplifier while permitting simultaneous transmission and reception at different frequencies through a device called diplexer. Moreover, the lower frequency band is normally used on the downlink to exploit the lower atmospheric attenuation or signal loss. Early satellites used frequencies between 2 and 8 GHz bands (C and X band). In recent years  $K_u$  bands at 14 and 16 GHz are used along with the C band.

The earth station in a satellite communication system consists of an antenna, a power amplifier and a low noise power receiver. Most stations are equipped with telemetry tracking and command system. Antenna diameter ranges from 0.7 m for direct broadcast receive only application to as large as 30 m for large international gateway stations.

## 12.7 Modems

As has already been said, analog signal is coded to form digital signal using pulse code modulation technique. Therefore, either in digital or analog communications square wave having wide frequency spectrum is transmitted in most of the communication systems. The problems of attenuation, delay distortions and noise limit the performances of communications. To avoid these problems, especially in telephone line, ac signalling is used. A continuous tone called carrier in the range of 1 to 2 kHz range is used. The variation in amplitude, phase and the frequency of the carrier wave is used to transmit signal. A device that accepts a serial stream of bits as input and produces a modulated carrier (or vice versa) is called modem, a short term of modulator-demodulator. The modem is in general inserted between a computer and the analog telephone system. Modems can be operated at 9.6 kb/s to 56 kb/s (as of today). Many modems now have compression and data correction code built into the system. The major advantage of increased data rate is offered by these features.

The interface between computer and the modem must specify functional and procedural standards. This is done by an internationally accepted EIA (Electronic Industries Association) standard called RS232C or its successor RS449. In this standard the computer is called the DTE (Data Terminal Equipment) and the modem is called DCE (Data Circuit-terminating Equipment). The connector has 25 pins; top row numbered as 1 to 13 and the bottom row 14 to 25. Figure 12.9 shows the 9 pins that are nearly always used in most of the applications. When the terminal or the computer is powered, it sets data-terminal-ready pin (pin 20) as



logical 1. When the modem is powered the data-set-ready pin (pin 6) is set to logical 1. When the modem detects a carrier on the telephone line, it asserts a logical 1 in carrier-detect pin (pin 8). Request-to-send (pin 4) then indicates that the terminal wants to send

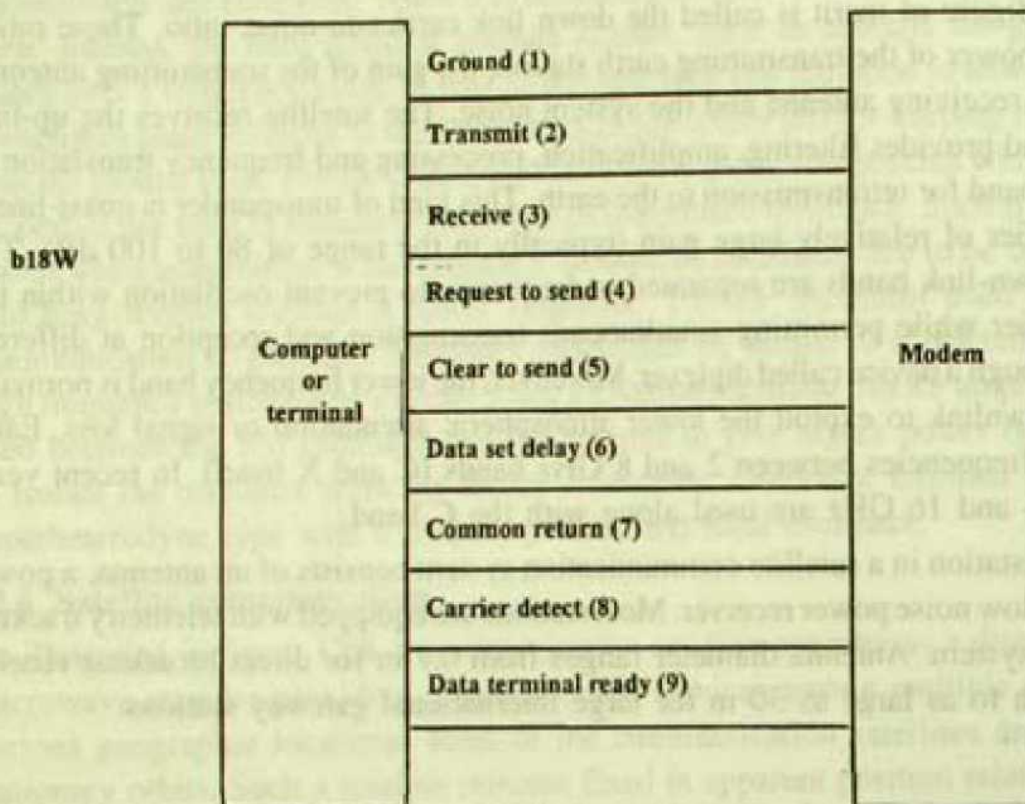


Fig.12.9 Major connection of RS 232C

data. Clear-to-send (pin 5) means that the modem is prepared to accept data. Data are then transmitted by the transmit-circuit via pin 2 and are received by the receive-circuit (pin 3). Other circuits and pins are provided for selecting the data rate, testing the modem, clocking the data and detecting ringing signals. These pins are not generally used.

## 12.8 Internet

A collection of interconnected network is called the internet. The interconnection is done by using computing machines called gateway which provide necessary translation between widely varying network in terms of hardware and software. The ability to connect multiple networks together in a seamless way is governed by a common protocol (rules) known as the transmission control protocol/internet protocol (TCP/IP). Therefore, a machine in the internet if it runs TCP/IP has an IP address and has the ability to send and receive data packets from other machines. Traditionally, the internet has four main applications:—

(a) Email: an electronic mail facility to compose, send and receive messages. (b) News: a specialized forum in which users with a common interest can exchange messages, (c) Remote log-in: using specific programme, users anywhere on the internet can log into any other machine on which they have an access. (d) File transfer: again using specialized programme it is possible to copy file from one machine on the internet to another. The





## SECTION-IV

### Computer

#### Chapter 13

#### Computer Hardware

##### 13.1 Introduction

The term “computer” originates from the word computation which means calculation. So a computer is a machine which can calculate with high speed. Computers available now-a-days can not only calculate but can also analyse data, take decisions and control external processes.

The best-known early attempt to make calculating machines was made by Charles Babbage. Starting from Babbage’s machine the computer has undergone a sea change with respect to speed, processing power, decision making capabilities etc.

Evolutionary phases of the computers are referred to as generations. Till date computers are classified into five generations.

- i) First generation : Early computers built with vacuum tubes fall in this category.
- ii) Second generation : In this category, transistors replaced vacuum tubes.
- iii) Third generation : In this category Integrated Circuit (IC) replaced transistors.
- iv) Fourth generation : With the use of V.L.S.I. IC (Very Large Scale Integration Integrated Circuit) chips in place of IC, computers of this generation came into existence.
- v) Fifth generation : Computers with some sort of intelligence belong to this generation. Computers of this generation are yet to be developed. Computer upto fourth generation works on Data / Logic Information Processing System (DIPS/LIPS), but the fifth generation computer would work on Knowledge Information Processing System (KIPS).

We see that with each generation, both cost and size are reduced drastically but with increased speed and processing power.

##### *Classification of computers*

Computers can be classified into three types :

- i) Analog computer.
- ii) Digital computer.
- iii) Hybrid computer.



An analog computer works on continuously variable inputs (e.g. electrical voltages), while a digital computer works on digital inputs and process digital data (having discrete values like 0,1 etc). Hybrid computer is partly analog and partly digital. It possesses some properties of analog and some properties of digital computer.

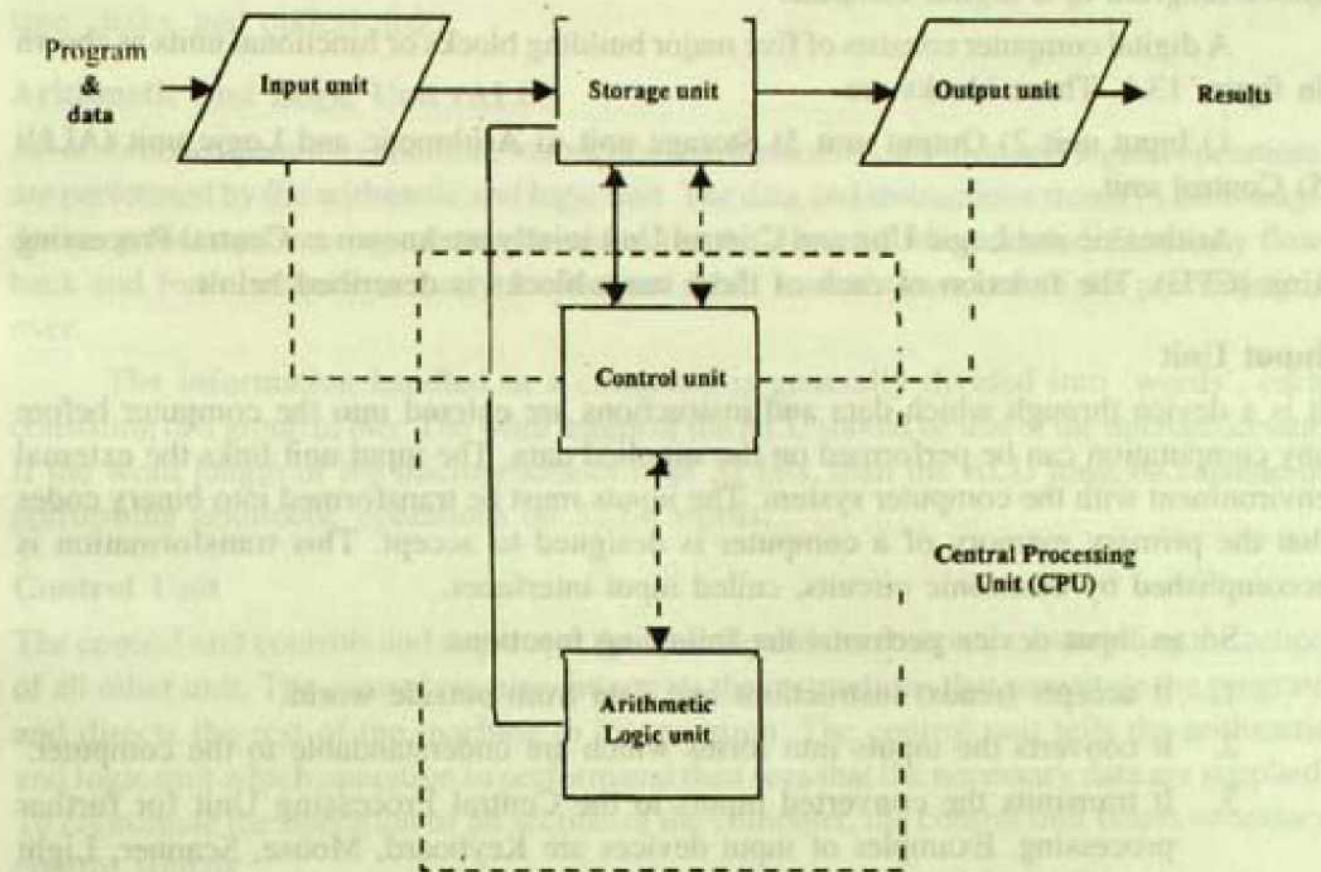


Fig. 13.1 Block Diagram of a Computer System

We shall confine our discussions only to digital computers. The entire range of digital computers can be subdivided into the following types :

- i) Micro computer ii) Mini computer iii) Main frame computer iv) Super computer.

The micro computers are the smallest in size and the speed and data handling capacity of earlier models were the least of all. But with the development of intel pentium-III/ AMD-K7 processor, speed and data handling capacity of a micro computer can match with that of a main frame computer. The micro computer was first developed by IBM Corporation and is named "Personal Computer" (PC). Micro computers are of three types :

- i) Personal Computer (PC)
- ii) Personal Computer with extended technology (PC/XT)
- iii) Personal computer with advanced technology (PC/AT)

Normally no hard disk is present in a PC and so to work with a PC, floppy diskettes have to be used to supply instructions and data. PC/XT and PC/AT both possess hard disk.



It is customary to name PC by the number of the microprocessor used namely 286, 386, 486, P, P-II, P-III etc.

### 13.2 Basic building blocks

#### *Block diagram of a digital computer*

A digital computer consists of five major building blocks or functional units as shown in figure 13.1. These blocks are

- 1) Input unit 2) Output unit 3) Storage unit 4) Arithmetic and Logic unit (ALU)
- 5) Control unit.

Arithmetic and Logic Unit and Control Unit jointly are known as Central Processing Unit (CPU). The function of each of these main blocks is described below.

#### **Input Unit**

It is a device through which data and instructions are entered into the computer before any computation can be performed on the supplied data. The input unit links the external environment with the computer system. The inputs must be transformed into binary codes that the primary memory of a computer is designed to accept. This transformation is accomplished by electronic circuits, called input interfaces.

So an input device performs the following functions :

1. It accepts (reads) instructions and data from outside world.
2. It converts the inputs into forms which are understandable to the computer.
3. It transmits the converted inputs to the Central Processing Unit for further processing. Examples of input devices are Keyboard, Mouse, Scanner, Light Pen, Floppy Disks, etc.

#### **Output Unit**

This unit receives processed data and results from the Central Processing Unit and supplies them to the outside world and so it links the computer system to the outside world. Before getting output results, which are in binary form, they must be converted to readable form. This conversion is performed with the help of electronic circuits called output interfaces.

In short, the output unit performs the following tasks :

1. It accepts the results from the computer in binary form.
2. It converts these results coded in binary to readable form.
3. It supplies the converted results to the outside world.

Examples of output devices are Monitor(VDU), Printer, Plotter, Floppy Disks, etc. Input and output devices are jointly known as peripheral devices.

#### **Storage Unit**

Storage unit or memory section of the computer consists of the devices to store the information that are entered into the computer through the input unit which will be used during



computation. The memory is also used to hold both intermediate and final results as the computer proceeds through the program. Memory devices are constructed so that the control unit can obtain any information residing in the memory. The time required to obtain information varies and is determined by the type of device used to store the information. Common storage devices are semiconductor based integrated-circuit memories, magnetic tape, disks, and optical disks.

### **Arithmetic and Logic Unit (ALU)**

All arithmetic operations addition, subtraction, multiplication, division and logical operations are performed by the arithmetic and logic unit. The data and instructions stored in the storage device prior to processing are transferred to the ALU as and when needed. Data may flow back and forth between memory section and ALU several times before the processing is over.

The information handled in a computer is generally divided into 'words', each consisting of a group of bits. The word length of the ALU should be that of the microprocessor. If the word length of the microprocessor is of 32 bits, then the ALU must be capable of performing arithmetic operations on 32-bit words.

### **Control Unit**

The control unit controls and sequences the operation of the computer, controlling the action of all other unit. The control circuitry interprets the instructions that constitute the program and directs the rest of the machine in its operation. The control unit tells the arithmetic and logic unit which operation to perform and then sees that the necessary data are supplied. To coordinate the operation of all section of the computer, the control unit issues necessary control signals.

## **13.3 Central Processing Unit**

The ALU and the control unit constitute the Central Processing Unit (CPU) of the computer. The CPU is the brain of the computer. The CPU also contains accumulators and general and special purpose registers which store data, intermediate results, memory address etc. during the execution of a program. The electronic circuitry and other physical components are known as the hardware of a computer system. Design of computer hardware deals with three concepts namely, computer organisation, computer design and computer architecture.

Computer organisation refers to how the hardware components operate and their interconnection to form the computer system. Assuming the various components to be in place, the task is to investigate the organisational structure to verify that the components perform the intended functions.

Computer design is concerned with the design of the computer from hardware standpoint. The design starts after the formulation of the specification is done. Computer design takes into account the type of hardware to be used and how they have to be connected. Computer design is also known as system implementation. Computer architecture is concerned with the structure and behaviour of the computer as seen by an user. It refers



to the information formats, the instruction set, and memory addressing techniques. Specifications of the various functional modules and structuring them together into a computer system is the subject of architectural design.

### 13.4 Storage System

A storage system is a device for storing data and instructions. Every storage unit of a computer system has the following characteristics.

1. *Access time* : This refers to the time required to locate and retrieve stored data from the storage unit in response to the instructions of a program.
  2. *Storage capacity* : This represents the volume of data that can be stored in the storage unit.
  3. *Cost per bits of storage* : This represents the cost to be incurred per bit of storage.
- On the basis of these three characteristics, storage units are basically of two types — Primary and Secondary.

#### Primary Storage

The primary storage is the main memory of a computer. Primary storage units have faster access time, smaller capacity and high cost per bit of storage. The main memory unit of a computer is made up of several small storage areas called locations or cells. Each of the locations can store a fixed number of bits. This fixed number of bits is the word length of primary memory of the computer. The primary memory is measured in terms of  $N = 2^{10} = 1024$  or 1 k byte. Each word or location has a built-in and unique number assigned to it and is called the address of the location. A particular location in the memory unit is identified by this unique number. Each location can hold either a data or an instruction. Irrespective of the contents, the address of the location remains the same. The organisation of primary memory is shown in fig. 13.2 given below

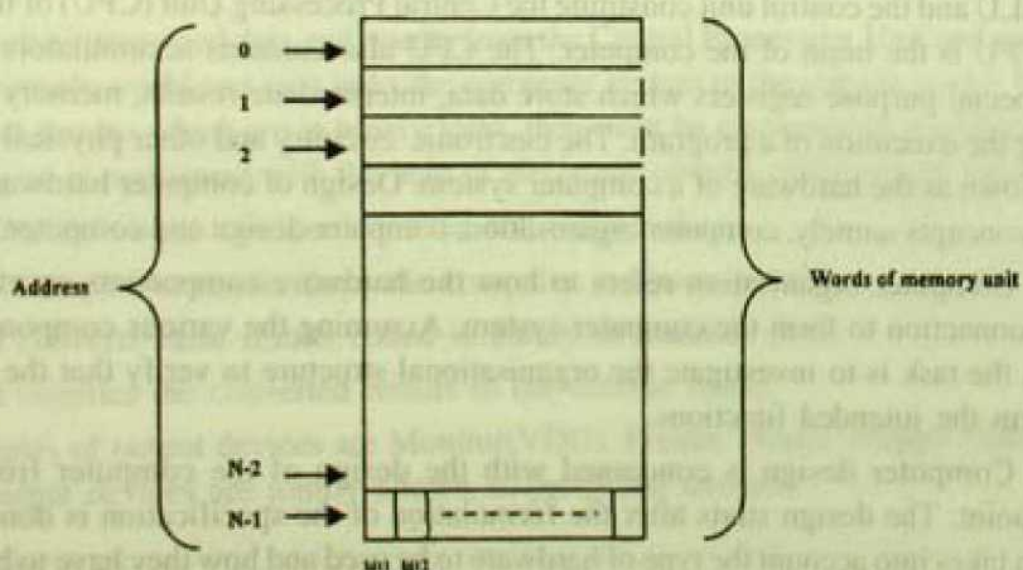


Fig. 13.2 Memory organisation



Normally addresses start at 0 and the highest address is equal to  $(N-1)$ , where  $N$  is the number of words that can be stored in the memory unit.

When the word at an address is read, it is not erased but word at that address is changed only when a new word is written or stored at that address. Capacity of memory is expressed in bytes (1 byte = 8 bits).

Now-a-days PC with "primary memory of several hundreds megabytes are available.

Two major types of primary memory are RAM and ROM.

### **Random-Access Memory (RAM)**

In random-access memory, the memory cells can be accessed for information transfer from any desired random location. That is, the process of locating a word in memory is the same and requires an equal amount of time no matter where the cells are located physically in memory organisation, thus the name "random access".

Communication between a memory and its environment is achieved through data input and output lines, address selection lines and control lines that specify the direction of transfer. The two operations that a random-access memory can perform are the write and read operations. A RAM is a general purpose device whose contents can be altered during computational process. RAM is a volatile memory, it is lost as soon as the main power is switched off.

### **Read Only Memory (ROM)**

The Read Only Memory is a permanent memory and only 'read' operation is possible; it has no capability of 'write' operation and hence the name "read only". The binary information stored in ROM are made permanent during the hardware preparation and the information stays within the unit even when the power is turned off and on.

Both the RAM and ROM chips are connected to the CPU through data and address buses (communication lines). The bidirectional data buses allow both way transfer of data between memory unit and the CPU.

*Cache Memory* : It is generally true that frequent references to memory is confined within a few localised areas in memory. So the total execution time of a program can be reduced if the active part of the program is placed in a fast small memory. Such a fast small memory is known as cache memory. This is placed between the CPU and main memory. The access time of cache memory is less than that of main memory by a factor of 5 to 10. Cache memory acts like a waiting room for the data which will be needed immediately.

### **Secondary Storage**

In a computer system large volume of data has to be stored; it is not possible to store them in the main memory due to high cost per byte of storing. Moreover, the most memory in RAM is not a permanent memory. This problem is overcome by using secondary storage



devices which are less expensive and are permanent, i.e. they are not lost, when the machine is switched off. The access time of secondary storage is much greater than the primary storage device (main memory). The secondary storage devices are used to store billions of bytes of data on a permanent basis which can be partially transferred to the main memory as and when required during processing.

Several secondary storage devices are in use, but the choice of a particular type depends on how the stored information needs to be accessed. Stored data are accessed in either of two ways — sequential or serial access and direct or random access.

In sequential access devices, stored data are accessed serially one by one starting from the first address and ending at the desired address. So in sequential access, access time depends on this location of the address. It is least for the first address and greatest for the last address. Sequential processing is suitable in such applications like preparation of monthly paybills, monthly electricity bills, etc where each address needs to be accessed in turn. Magnetic tape is one of the sequential storage devices.

### 13.5 Magnetic disks

In random access every information stored can be accessed randomly i.e. any location can be accessed directly and access time for all locations is approximately equal. Random access is suitable for such applications as banking service, reservation, etc. Magnetic disk and magnetic drum are the two most commonly used direct storage devices. Magnetic disks are of two types namely hard disk and floppy disk.

#### Hard Disk

A hard disk is a magnetic disk made of a thin metal plate, both sides of which are coated with a magnetic material. They are housed in dust free environment and are permanently attached to the unit assembly. A hard disk is a secondary storage device and can be used for both sequential and random access. Often several disks are stacked on one spindle. All disks rotate together at high speed. Bits are stored in the magnetised surface in spots along concentric circles called tracks. Tracks are commonly divided into sections called sectors. Minimum quantity of information which can be transferred is a sector, subdivision of a disk surface is shown in Fig. 13.3. In a disk pack data are not stored on the upper surface of the top disk and the lower surface of lower disk. In some units a single read/write head is used for each disk surface. The read / write head is movable along a radius. A disk system is addressed by address bits that specify the disk number, the disk surface, the sector number and the track within the sector. Bits are recorded on track either with equal density or with variable density. The information stored on a disk can be read many times without affecting the stored data. Writing of new data erases the previous data. Storage capacity of a hard disk depends on the number of disks in the pack and is very high. Disk with capacity of several Giga Bytes (GB) is available now-a-days. Average access time of a disk is usually between 10 and 100 milliseconds.



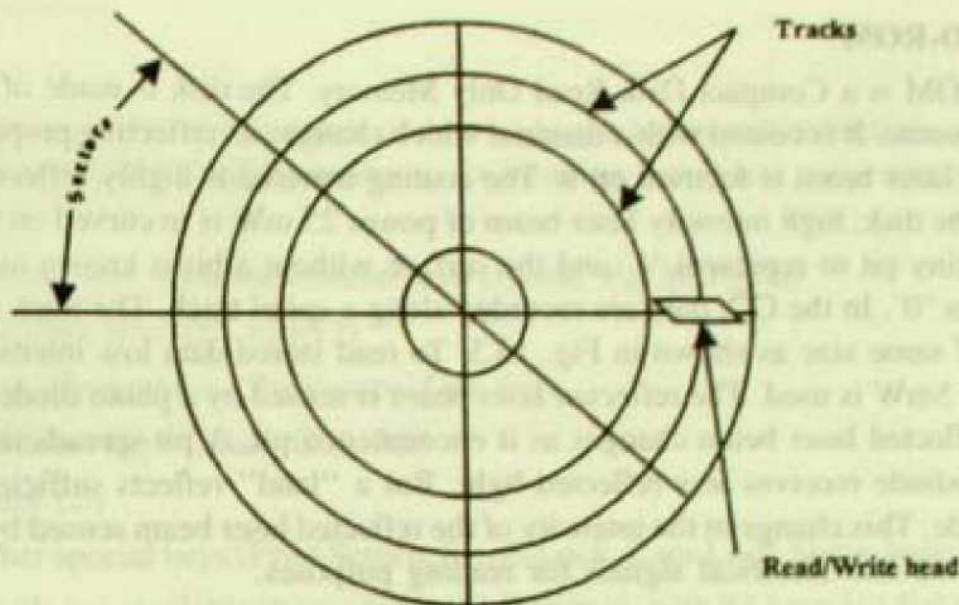


Fig. 13.3 Magnetic disk

### Floppy Disk

A floppy disk is a magnetic disk and is used as a secondary storage device. Floppy disks are also known as diskettes or floppies. A floppy disk is made of flexible plastic (Mylar) which is coated with magnetic material (iron oxide). The disk is enclosed in a square plastic jacket which protects the disk from dust. Data are stored in the magnetised surface in spots along concentric circles called tracks. The tracks are commonly divided into sections called sectors. These disks are used with a floppy disk drive and are removable. There are two sizes commonly used, with diameter 5.25- and 3.5 inches. The protective cover of 5.25 inch floppy is soft, while that of 3.5 inch floppy is hard. The storage capacity of the smaller size is 1.44 MB and of the larger one is 1.2 MB (360 KB for low density 5.25 inch floppies). Read /write operation is done on the disk through an aperture in the jacket. The read / write head makes direct contact with surface of the disk and so the floppies get worn with constant use.

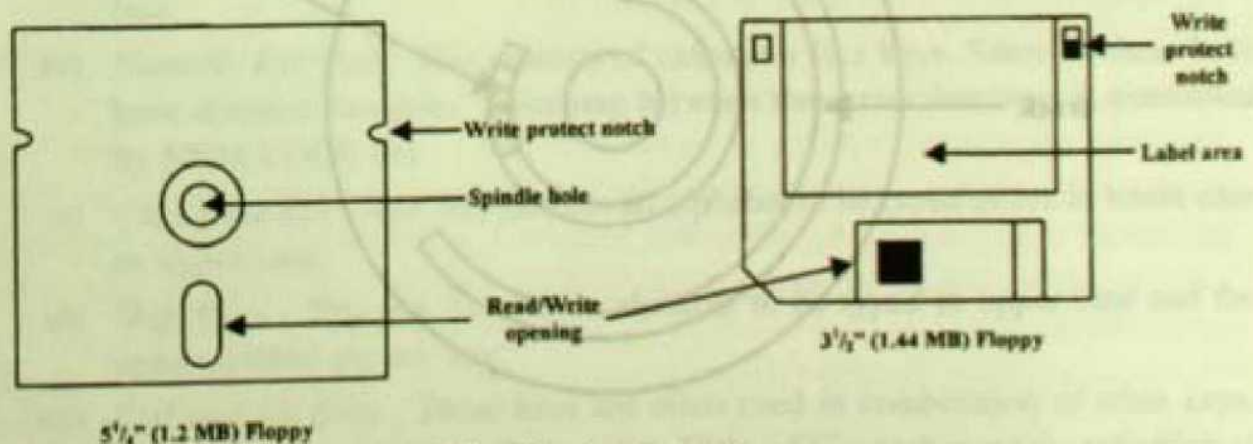


Fig. 13.4 Floppy disk



### 13.6 CD-ROM

A CD-ROM is a Compact Disk-Read Only Memory. The disk is made of resin, such as polycarbonate. It is coated with a material which changes its reflecting property when high intensity laser beam is focused on it. The coating material is highly reflective. To record data on the disk, high intensity laser beam of power 25 mW is focused on the disk which forms a tiny pit to represent '1' and the surface without a pit is known as 'Land' and it represents '0'. In the CD, data are recorded along a spiral track. The track is divided into blocks of same size as shown in Fig. 13.5. To read stored data low intensity laser beam of power 5mW is used. The reflected laser beam is sensed by a photo diode. The intensity of the reflected laser beam changes as it encounters a pit. A pit spreads the light so that the photodiode receives less reflected light. But a "land" reflects sufficient light to the photodiode. This change in the intensity of the reflected laser beam sensed by a photodiode is converted into electrical signals for reading purposes.

If the coating of the special material is a thin film, a hole is formed when the strong laser beam falls on it and for a thick film, a bubble is produced for commercial production, first a master disk is produced. The master disk is used for mass production of CD-ROM. Copies of CD-ROMs are produced by moulding the master disk with a special plastic. The plastic copy of the data is then covered with a clear plastic layer and backed with a coating of reflecting material. To read data a read head passes over the track at a constant linear velocity. The CD has to be inserted into the CD-ROM drive provided with the computer. Data on a CD-ROM is written only once and can be read again and again. A CD-ROM can store data to the tune of several hundred megabytes. It has a long life time since there is practically no wear and tear. Now-a-days CD-ROMs are used extensively for storing system software, application software and games.

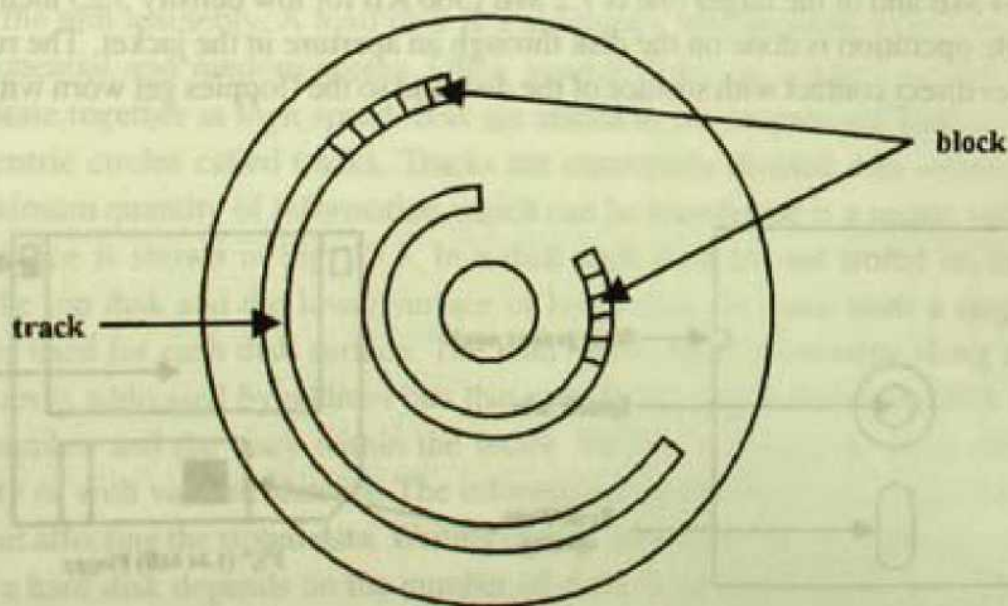


Fig. 13.5 Spiral track of a CD-ROM



## 13.7 Input Devices

### Keyboard

Keyboard is an important input device, through which data and instructions are entered into the computer. A keyboard is similar to a typewriter, but it has some additional keys given below :

- Arrows and other movement keys (Home, PgUp, PgDn)
- Function keys (F1 to F12)
- Text editing keys (Backspace, Del, Ins)
- Modifier keys (Shift, Ctrl, Alt)
- Enter (↵)
- Other special keys (Print Screen, Scroll Lock, Caps Lock Status indicator LED.)

Keyboards are available in two forms (i) Standard with 84 keys (ii) Enhanced with 101 or more keys. The enhanced keyboards are most popular. They have some additional keys for multimedia and internet. The functions of the different keys are described below :

- i) *Typewriter keys* : These keys include letters, numbers and punctuation symbols.
- ii) *Function keys* : These are the keys labelled F1 to F12. These keys carry out different functions depending on the software used.
- iii) *Cursor Control Keys* : These keys are marked ←, →, ↓, ↑ and are called the Left, Right, Down and Up Arrow keys respectively. The cursor keys are used to move the cursor left, right, down or up around the screen, one line or one character at a time. These are four other cursor control keys, just next arrow keys. These keys are labelled as Home, End, PgUp and PgDn. PgUp key is used to move to the previous screen or page of the document and PgDn key is used to move to the next screen or page of the document. Home key is used to move the cursor to the top of the document or at the beginning of a line while the End key is used to move the cursor to the end of the document or end of the line.
- iv) *Numeric Key Pad* : This consists of calculator like keys. Some of these keys have doubled functions. Switching between these two functions is controlled by NUM LOCK key.
- v) *Caps Lock Key* : This key enables an alphabet to be typed either in lower case or upper case.
- vi) *Shift Keys* : This key enables an alphabet to be typed in upper case and the upper symbol on any key.
- vii) *Ctrl and Alt Keys* : These keys are often used in combination of other keys, to produce special effect. By pressing Ctrl and C simultaneously execution of



current task is abandoned and the machine returns to the DOS prompt. The machine is automatically restarted if Ctrl, Alt and Del keys are pressed simultaneously.

- viii) *Enter/Return Key* : After typing a command this key is pressed to enter the command and then the execution will start. By pressing this key a new line or a new paragraph begins.
- ix) *Pause* : Pressing this key stops scrolling and helps to read the screen.
- x) *Tab* : The Tab key moves the cursor to a present position along a line. In some software, the Tab key allows to move from one option to another option in a menu.
- xi) *Esc* : This key is used to Cancel or Ignore the entry or command that has been just entered.
- xii) *Print Screen* : This key is used for printing the text of the screen on a printer.
- xiii) *Delete* : Pressing this key erases the character on which the cursor is presently on.
- xiv) *Backspace* : Pressing this key, erases the character on the left of the character when the cursor is presently on.
- xv) *Windows Keys* : There are two more keys now available in the latest keyboards, called Windows Keys, identified by the windows icons on them. They are used to PgUp the start button. There is another key which is used to bring up right click menu. This key is identified by the mouse cursor and menu icon on it.
- xvi) *LED Status Indicator* : The on state of the LEDS indicate the status of the Caps Lock and Menu Lock Keys.

### *Mouse*

A mouse is a hand-held pointing device to control the operation of a computer. The mouse is normally used as an input device to a computer with Windows operating system. The mouse is used to select specific options from the menu. A mouse is a rectangular box with a rubber ball at the bottom and two or three buttons on the top surface. The mouse cursor is an arrow whose movement is controlled by the mouse. The left button is used frequently for selecting options while the right button is used for special effect. To select an item in the menu, the mouse cursor is pointed to that icon by moving the mouse on the mouse pad and then the left button is clicked. Clicking the right button will do the same function as done by double clicking the left button. The middle button is used to scroll the screen.

The hourglass symbol along with the arrow indicates that processing is taking place. After the completion of processing the hourglass vanishes.



## 13.8 Output Devices

### *Printer*

Printer is a most commonly used output device for hard copy of the result of computation. As the result produced is in binary form, it must be converted into readable form (English like). This task is accomplished by the output interface. After this conversion, the result goes to the output device. Two types of printers are the character and the line printers. Character printers print one character at a time. Line printers print several characters on a single line so that they appear to be printed at the same time.

Printers can also be classified as impact and non-impact printers depending on the techniques that are used to generate the characters. In impact printers, the printing element hammers or pins the printing medium directly in order to generate the character. In non-impact printer thermal, chemical or electrostatic methods are used to print a character rather than direct impact. Another way to classify printers is by using character formation techniques :

Matrix printers utilise dots to form characters. Character printers use completely formed characters.

### *Dot-Matrix Printers :*

These printers print each character as a pattern of dots. The print head comprises a matrix of tiny needles, typically seven rows with nine needles in each, a  $9 \times 7$  matrix, which hammers out characters in the form of patterns of tiny dots. The shape of each character i.e. the dot pattern is determined by the information held electronically in printer. Character font of a dot-matrix printer is not fixed and so they can print any shape of character in different sizes and have the ability to print charts and graphs.

### *Ink-Jet Printers :*

Ink-Jet printer are of non-impact type. The ink-jet printer prints characters by spraying small drops of ink on the printing medium. Special type of ink having high iron content is used. Droplets of ink are electrically charged after leaving the nozzle. The droplets are then guided to the proper position in the printing medium by electrically charged deflection plates, generating the character. Print quality of ink-jet is much better than that of dot-matrix printers. Colour printing is also possible with inkjet printers.

### *Laser Printers :*

Another non-impact type printer is a laser printer. A laser printer prints one page at a time. It uses laser beam to produce an image on a photo sensitive drum. The computer controls the movement of the laser beam across the surface of the drum. The spots in the surface of the drum exposed to the laser beam becomes charged and attract ink powder. Thereafter the drum transfer the ink-powder to a paper where the ink powder is permanently fused giving a hard copy of the document to be printed. The drum is then discharged and cleaned and is then ready for printing the next page (DTP). Printing quality of a laser printer



is very excellent and is normally used for desk top publishing (DTP). A laser printer can print 300 pages per minute.

### **Monitor**

A monitor is an output unit which is used to display what is typed in the keyboard and also the output of a program. The core of a monitor is a cathode ray tube (CRT) or liquid crystal display (LCD). The most common type of monitors used are the VGA (Video Graphic Array) monitor which can display both text as well as graphics. They are available in monochrome (B/W) as well in colour type.

### **13.9 Software Concept**

A set of logically related instructions written in computer language in a specific sequence to accomplish a task is called a program. The program controls the operation of the computer and it performs the task wanted by the programmer. A set of program procedures and associated documentations is generally referred to as software. Without a software a computer is nothing but a junk.

#### *Classification of Software*

Softwares are classified into the following categories :

- i) System Software
- ii) Application Software
- iii) Utility programs
- iv) Packages.

#### **System Software :**

System Software are sets of one or more programs that are designed to control the operation of a computer in more effective and efficient manner. A System Software supports the following operations :

- a) Running of other software
- b) Communicating with peripheral devices
- c) Development of other types of softwares
- d) Monitoring the use of various hardware resources.

A few types of system softwares are

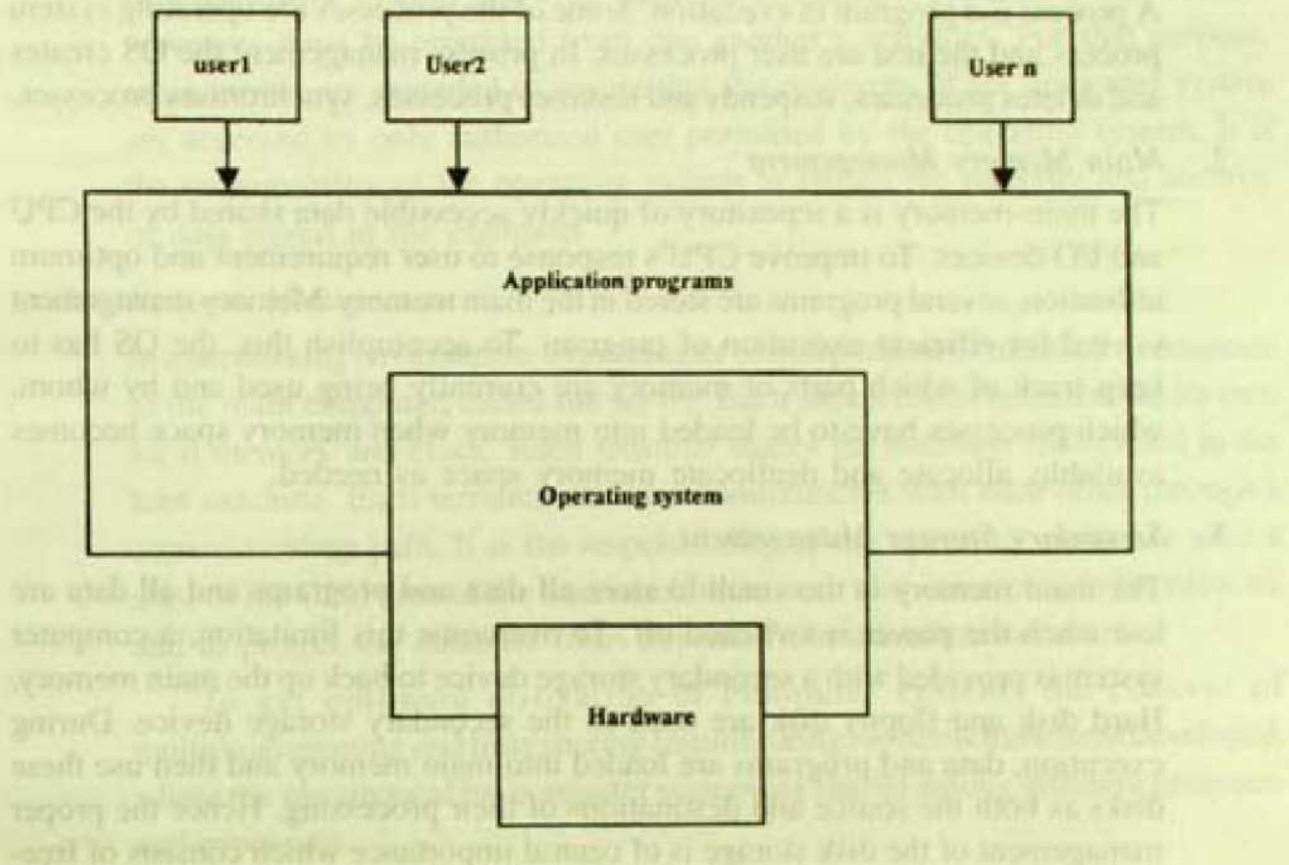
- i) Operating Systems
- ii) Assemblers
- iii) Compilers
- iv) Interpreters



## Operating Systems :

An Operating System (OS) is a master control program that runs the computer and acts as a scheduler. It acts as an intermediary between a user and the hardware. It controls the flow of signals from CPU to various parts of the computer. The operating system provides an environment for executing programs efficiently and effectively.

A computer system can be divided into four components : hardware, operating system, application programs and users.



**Fig. 13.6** Abstract view of the components of a computer system.

### 13.10 Operating System Concept

The interface between the operating system and the user programs is defined by the set of extended instructions that the operating system provides. This set of extended instructions is known as systems calls. The system calls create, delete and use various software objects managed by the operating system. The most important of these are processes and files.

#### *Processes*

Process is a key concept in all operating systems. A process is basically a program in execution. It consists of the program, data and all other information needed to run the program. Creation and termination of a process is handled by process management system calls.



## *Files*

Another important system calls is related to the file system. System calls are needed to create files, delete files, read files and write files. In order to read a file it must be opened and after the file has been read, it must be closed. System calls are provided to do these tasks.

The operating system (OS) acts as resource manager. The following managements come under this head :

1. *Process Management :*

A process is a program in execution. Some of the processes are operating system process and the rest are user processes. In process management the OS creates and deletes processes, suspends and resumes processes, synchronises processes.

2. *Main Memory Management :*

The main-memory is a repository of quickly accessible data shared by the CPU and I/O devices. To improve CPU's response to user requirement and optimum utilisation, several programs are stored in the main memory. Memory management is vital for efficient execution of program. To accomplish this, the OS has to keep track of which parts of memory are currently being used and by whom, which processes have to be loaded into memory when memory space becomes available, allocate and deallocate memory space as needed.

3. *Secondary Storage Management :*

The main memory is too small to store all data and programs and all data are lost when the power is switched off. To overcome this limitation, a computer system is provided with a secondary storage device to back up the main memory. Hard disk and floppy disk are used as the secondary storage device. During execution, data and programs are loaded into main memory and then use these disks as both the source and destinations of their processing. Hence the proper management of the disk storage is of central importance which consists of free-space management, storage allocation and disk scheduling.

4. *I/O System Management :*

It consists of assigning I/O system to different programs that are being executed.

5. *File Management :*

A file is a collection of related information consisting of data and programs. A file consists of bits, bytes or records whose meaning are defined by their creators. The operating system implements the abstract concept of a file by managing disks or tapes and the devices which control them. Files are normally organised into directories and there may be multiple users accessing the files. The operating system is responsible for creation and deletion of files and directories, manipulation of files and directories, mapping of files on secondary storage devices, back up of files on stable storage media.



#### 6. *Device Management :*

During execution of a program, additional resources, such as more memory, I/O devices, files etc may be required for completion of the execution of the program. If resources are available then the OS will grant the request, otherwise the program has to wait until the resources requested are available. This is pertinent with multiuser system.

#### 7. *Management of System Protection*

In a multiuser system allowing multiple concurrent processes, this various processes must be protected from one another's activities. For that purpose, mechanisms are provided to ensure that the resources of the computer system are accessed by only authorised user permitted by the operating system. It is the responsibility of the operating system to ensure the integrity and security of data stored in the computer.

#### 8. *Network Management :*

In a networking environment large number of independent terminals are connected to the main computer, called the server. Each independent terminal has its own local memory and clock. Each terminal shares the database maintained in the host machine. Each terminal can also communicate with each other through a communicating path. It is the responsibility of the operating system to allot a slice of the CPU time of the host machine to each user connected to the network and to protect the database from mis-use from unauthorised users.

To get optimum utilisation of computer systems the concept of multiprogramming and time sharing (multitasking) systems have been developed, where the resources of the computer system are shared among different programs and processes.

### **MULTIPROGRAMMING**

The term multiprogramming refers to interleaved execution of different independent jobs by the same computer. This concept has been developed to minimise the CPU's idle time and to keep it always busy executing part of a program. In multiprogramming environment a number of jobs are batched and kept ready. The Operating System keeps these jobs in the memory at a time. The OS picks and begins to execute one of these jobs stored in the memory. Occasionally this job may have to wait for some task. During this time the OS switches and executes another job. When the second job needs to wait, the OS switches to the third job. This switching from one job to the next job continues until the last job is reached. In the time in which the CPU switches from the first to last job, the first job may finish its waiting and gets the attention of the CPU. So as long as there is some job, to execute, the CPU will never be idle.



The job control program of the OS is given charge of executing these programs one after another. To accomplish sequential execution of the jobs, instructions are written in a special language, known as Job Control Language (JCL). These instructions tell the OS the names of the jobs, the files required and the devices to be used.

### *TIME-SHARING*

Multiprogramming systems suffer from two difficulties : i) Multi-programming execution takes place in multisteps in which subsequent steps may depend on the result of earlier ones. As the user cannot interact with the job it is executing, it is difficult to take care of unwanted result. ii) It is not possible to modify a program to study its behaviour, as the program is executing. Time sharing is the extension of multiprogramming where both of the above difficulties have been overcome. The term time sharing refers to the computer system that has a number of independent, relatively slow speed, on-line terminals connected simultaneously to the main computer of the system. Each terminal provides a direct access to the CPU and shares the resources in a time-dependent fashion.

The basic idea behind the time sharing system is to allot a slice of the CPU time to each user. During his time slice, an user gets the attention of the CPU and a portion of his program is executed. The CPU switches continuously from one user to another user and a portion of the program is executed, until the job is completed. The speed of the CPU is so fast and the switching time is so small (about 20 millisecond) that each user has the feeling that he is the lone user and processing at his terminal appears to be continuous. Although it may seem that several users are using the same computer simultaneously, but in actual case, only one program is under the control of the CPU at a time and only one instruction is executed at a time. In time sharing environment, the CPU does not wait idle.

### *UNIX*

The operating system UNIX was designed to be a time-sharing and multiprogramming system. UNIX consists of two separate parts, the Kernel and the systems programs. The layer below the system call interface and above the physical hardware is the Kernel. The Kernel provides the file system, CPU scheduling, memory management, and other operating system functions through system calls. Systems programs use the Kernel supported system calls to provide useful functions, such as compilation and file manipulation.

System calls define the programmer interface to UNIX. The set of systems programs commonly available defines the user interface. The programmer and user interface define the context that the Kernel must support.

Both user-written and systems programs are executed by a command interpreter, which is a user process. It is called shell as it surrounds the Kernel. The shell initialises itself, and then types a prompt character often a dollar sign, on the screen, where the command is typed. The shell is the primary interface between user and the operating system. The shell is not part of the Operating System and can be replaced by another without affecting the Kernel. Writing a new shell would greatly change the user view. Replacing the keyboard



oriental shell by screen oriented shell leads to the development of X-Windows system, which has become a standard.

System calls for Unix are roughly grouped into three categories : file manipulation, process control and information manipulation. Devices in Unix are treated as file and so same system calls can support both files and devices.

## *DOS*

Disk operating system (DOS) is the most popular operating system. DOS is neither multiprogramming nor time-sharing operating system. DOS is a single-processor operating system and is used in console mode of operation. DOS is available in two forms — Microsoft Disk Operating System (MS-DOS) and Personal Computer Disk Operating System (PC-DOS) of IBM corporation. MS-DOS is widely used. MS-DOS is structured in three layers, as follows :

1. The BIOS (Basic Input Output System)
2. The Kernel
3. The shell, command . Com.

The BIOS is a collection of low-level device drivers that serve to isolate MS-DOS from the details of the hardware. The BIOS contains calls to read and write from the disk, read a character from the keyboard and to write a character to the screen. A part of BIOS is located in ROM. The rest of the BIOS procedures not contained in ROM is held in a hidden file called *io.sys*. It is loaded immediately after the computer is booted and provides a procedure call interface to BIOS, so the Kernel can access BIOS services by making procedure calls to *io.sys*. The existence of *io.sys* further isolates the Kernel from hardware details.

The Kernel is contained in another hidden file, *msdos.sys* and contains the machine-independent part of the operating system. It handles process management, memory management and file system, as well as the interpretation of system calls.

The third part is the shell, *command.com* which is, although not a part of the operating system, uses many features of the operating system. It is the primary interface between the user and the operating system. The shell interprets the command typed by the keyboard and causes the Kernel to perform the desired task.

When MS-DOS is loaded properly, its readiness to accept command is indicated by a symbol on the screen called DOS prompt. The nature of the DOS prompt is

either *C:\>* or *C >*

A user command is typed at the DOS prompt.

## *MS-DOS File System*

MS-DOS file names have a basic part of 8 characters followed by an optional extension of maximum 3 characters. The extension part always starts with a dot. Only upper case characters are used and no distinction is made between upper case and lower case characters.



Only alphabetic, numeric and a few other characters (e.g. the underscore) can be used in the file names. Some common MS-DOS file extensions are listed below :

- bat — Batch file
- com — Single segment executable binary file
- doc — Documentation file
- exe — Executable binary file
- obj — Object file.
- sys — Device driver or other system file
- txt — Text file.
- for — FORTRAN source file
- cbl — COBOL source file

The functions performed by DOS can be classified broadly into two categories. One category of jobs are performed automatically and the second category is performed when asked for. The entire set of DOS commands are classified as internal commands and external commands. Some internal commands are dir, del, rename, copy, type etc. These command are available from the resident part of the operating system. Some external commands are format, diskcopy, etc. These are in the transient part of the operating system and are brought into the main memory temporarily when needed.

## WINDOWS

Windows is an user friendly multitasking operating system that works on Graphical User Interface (GUI). The graphical interface allows user to select files, programs or command by pointing to pictorial representation on the screen, called icons by means of a pointing device, such as a mouse. The three most common ways to use a mouse include pointing, clicking or double clicking and dragging.

### 13.11 Computer Programming

A program is a set of logically related instructions in a specific sequence to accomplish a given task. An instruction is a command given to the computer to perform a specified operation on the data supplied. Writing a program for a computer consists of specifying a sequence of machine instructions. Machine instructions inside the computer form a binary pattern, which is difficult to understand. So it is necessary to write programs using the more familiar symbols of the alphanumeric character set. But a program written in alphanumeric character set is not intelligible to the CPU and hence to run program it is necessary to translate user-oriented symbolic program into binary programs (combination of binary bits 0 and 1) recognised by the CPU. A translator is a program that translates user-oriented symbolic program (source code) into machine language (object code) compatible with the microprocessor being used in the computer.

If the source code (source program) is in assembly language (low-level language), then the translator is called an assembler. If the source program is in a high-level language



(e.g. COBOL, FORTRAN, C, etc.) then the translator is called a compiler. The translator is an interpreter if the source program is the BASIC.

An interpreter reads one statement of a high-level language program, translates it into object code and executes it. Then it reads the next statement of the program, translate and executes. This continues until the last statement is executed. On the other hand, a compiler reads the entire high-level language program, checking it for any syntax error and then translate the entire program into object code. A linker generates an executable file in machine language from the object code, linking various parts of the program and assigning specific locations for the memory required DOS. The executable file (usually with extension .exe or .x in UNIX) can then be run to execute the program. An interpreter is a smaller program than a compiler and occupies less space in memory than that of a compiler program. A compiler is much faster than an interpreter.

### *Computer Languages*

The language in which a computer program is written is called a computer language. The set of rules of a computer language that must be strictly followed in writing a program is the "syntax rules" of the language. Computer languages can be classified broadly into the following categories.

- i) Machine language
- ii) Assembly language
- iii) High level language

The set of instructions written in binary codes, which is understood by a computer without a translating program, is called a machine language. Machine language program is machine dependent. A language which uses mnemonics is called an assembly language. English like symbolic code for each instruction is called a mnemonic. The complete set of mnemonics is called assembly language and the program written in mnemonics is called an assembly language program. Machine language and assembly language are microprocessor specific and both are considered as low-level languages. The computer does not understand the assembly language directly. A program which translates an assembly language program into a machine language program is called an assembler. The language, in which a program written is not microprocessor specific, is called a high-level language. A high level language is procedure-oriented rather than computer oriented in the sense that high-level language permits writing problem oriented programs. The instructions written in high-level language are called statements.

## Chapter 14

### Programming in Fortran

#### 14.1 Introduction to Fortran

The computer, as it is used ordinarily, has a large memory and can follow instructions blindly and absolutely faithfully but has no intelligence of its own. Hence to use the computer, one must instruct it much as one instructs a child to solve a problem. For example, suppose we want the computer to solve the set of equations

$$6x + 5y = 1$$

$$2x + 3y = -1$$

We cannot just give the computer these equations and expect that the computer will come up with the answer by itself. We must instruct the computer. The set of instructions are to be given in a properly planned step by step procedure. This step by step sequence of instructions to solve the problem is called an "algorithm". This is sometimes depicted in a pictorial representation called "flowchart".

The next step is that the computer must understand our instructions — so we have to use a language to instruct the computer. But we cannot use a natural language (like English, Bengali, etc.) which is unambiguous and context free. The computer acts like a slave having no intelligence; hence the language must be completely unambiguous and context free, and have specified rigid grammatical rules. A number of such high level languages have been developed. We will discuss one of them, viz. FORTRAN, which is most commonly used in physics research. There are several other languages, designed for specialised uses, for example, BASIC, ALGOL, COBOL, C, etc. The set of instructions written in a computer language is known as a "computer program".

Thus the steps in solving a given problem will be as follows :

- (1) Choose a suitable method for solving the problem.
- (2) Write the algorithm. It is instructive to draw a flowchart as well.
- (3) Write the program in a suitable computer language. This is called a source code file (say TEST. FOR).
- (4) Feed the instructions into the computer (called keying in), using the key board.
- (5) Ask the computer to check for grammatical (called "syntax") errors. This is done by the command

FORI TEST;



If there are errors, they should be corrected and this process is repeated until completely free from errors.

- (6) Next ask the computer to translate the program in the machine language by using the command e.g. PAS2. This command creates an object code (TEST.OBJ).
- (7) The computer then needs assignment of memory locations for each of the variables used in the program and how they are linked by the program. This is done by the command

LINK TEST;

This creates a file named TEST.EXE.

- (8) Ask the machine to execute the program by the command  
TEST.EXE

and supply the data, in the manner in which it was planned in the program. The machine will then execute the set of instructions and supply the output results in the manner prescribed in the program.

## Fortran Language

We will discuss only the rudiments of the FORTRAN language here. Many details and additional features are omitted. Interested readers can learn more from any standard book on this subject, e.g. "Computer Programming in FORTRAN 77" by V. Rajaraman or "FORTRAN IV programming" by McCracken.

The name FORTRAN comes from FORMula TRANslation. It has undergone a lot of changes since it was first conceived. The first popular standard version (1966) was FORTRAN IV and the updated standardised version of 1977 is called FORTRAN 77. These are the most popular versions used all over the world.

In a computer program, all FORTRAN statements can start from the 7th column (not from the 1st to 6th columns) and must not continue beyond 72nd column. If the statement is too long, it can be continued in the next line with a character in the 6th column, indicating continuation. FORTRAN statement numbers (see below) can appear in 1st to 5th columns. A "C" in the 1st column indicates a comment, which is ignored by the computer.

In the following the number 0 (zero) will be denoted by Ø. Also the symbol ⇐ will indicate pressing the "enter" key board.

### 14.2 Constants and Variable names

In writing a program, one needs fixed numbers called constants like 4, 3, 2.1, 5.753 etc. In FORTRAN two types of constants appear:

(1) Integer constants like Ø, 1, 2, 3, -7

(2) Real constants like 1.25, 3.71, -5.12, etc. These may be written in exponent form, e.g. 1.57E-3 is Ø.ØØ157 and 2.57E5 is 257ØØØ.Ø. A real constant with a whole number value should be followed by a decimal point (e.g., 125.0 or 125.).



## Variable names

A quantity which may vary during program execution is called a variable. The name can be any combination of one to six letters or digits; the first character must be a letter. Examples: A, B2, THETA, INDEX etc. There are two types:

(1) *Integer variables* are those which can take integral values only. The name must begin with one of the six letters I, J, K, L, M, N. Examples: INDEX, I, KOUNT, IA, etc.

(2) *Real variables* are those which can take real constant values. The name must begin with any letter other than I, J, K, L, M, N. Example: A, CAB, THETA, B2C1, AIJK, etc.

It is possible to use a variable name beginning with I, J, K, L, M, N as real variable by a TYPE statement at the beginning of the program:

REAL INT, JX

Similarly it is possible to use a variable name beginning with any letter other than I, J, K, L, M, N as integer variable by the TYPE declaration:

INTEGER COUNT

The computer assigns a memory location for each variable name and the contents of that location is the present value of the variable. This can be thought of as a box with the name of the variable as a label and the content of the box is the present value of the variable.

## Subscripted variables

Often one has to use an array of numbers with a single name. An example is the expression

$$c_i = \sum_{j=1}^{10} a_{ij} b_j \quad (i = 1,5)$$

The quantities b and c have one subscript each and a has two subscripts. These can be represented as subscripted FORTRAN variables. The subscripts must be integers and the maximum value that each subscript can take must be specified by a DIMENSION statement at the beginning of the program. For the example cited above, we need

DIMENSION A(5,10), B(10), C(5)

In the course of the program execution, the value of the subscript must be an integer  $\geq 0$  and  $\leq$  the value specified in the DIMENSION statement. We can also have subscripted integer variables, e.g.

DIMENSION IN(5), MATRIX(4,4)



### 14.3 Arithmetic operators and modes for expressions.

The arithmetic operators are:

Fortran operator symbols	Arithmetic operations
+	addition
-	subtraction
*	multiplication
/	division
**	exponentiation

Two FORTRAN variables or constants are connected by an operator to form an expression. Example

$$\begin{aligned} (a+b)^2 &\rightarrow (A+B)**2 \\ \frac{(xy-a)}{b} y^c &\rightarrow ((x*y-A)/B)*(y**c) \end{aligned}$$

We can use a number of brackets (only first bracket symbols). The number of left brackets must match with the number of right brackets. The computer starts evaluating the expression from the innermost bracket and then proceed outward as indicated by the last example.

If specific brackets are not used, the hierarchy of operations is: exponentiation (\*\*) *first*, multiplication (\*) and division (/) *second*, addition (+) and subtraction (-) *last*. Also expressions are evaluated from left to right: all exponentiations are done first; the expression is then scanned again from left to right for divisions and multiplications and executed in the order of their appearance. Finally all the additions and subtractions are done starting again from the left to the right. Example:

$$\begin{aligned} A**B/C+D**E*F-H/P*R+Q &\rightarrow A^B/C + D^E*F - H/P*R+Q \\ &\rightarrow \left(\frac{A^B}{C}\right) + (D^E)F - \left(\frac{H}{P}\right)R + Q \end{aligned}$$

Another operation is the assignment operation (denoted by =). The left side of the = sign must be a variable (integer or real, ordinary or subscripted) and the right side can be a FORTRAN constant, (integer or real) or a FORTRAN expression. For example:

$$B2C1 = 5.0$$

$$C = (A+B)**2$$

The computer evaluates the expression on the right and assigns the value to the variable on the left side. Thus in the first example, the variable B2C1 is assigned a value 5.0. In the second example, the computer picks up the values of the variables A and B (from the corresponding memory locations or "boxes", adds them up, then squares the number thus obtained and assigns this number to the variable C, i.e. returns this number to the memory location for the variable C. If a variable is not already assigned, its value is taken as 0.

Thus if we have

$$C = (A+B)**2$$

$$A = 2.0$$

$$B = 3.0$$

then the value of  $C$  will be  $0$  (since FORTRAN statements are executed in the order in which they are encountered). But if we have

$$A = 2.0$$

$$B = 3.0$$

$$C = (A+B)**2$$

then the value of  $C$  will be  $25.0$ . It is not desirable (at least at the beginning) to mix up real and integer variables in a given expression (unless an integer is used as an exponent) or on either side of  $=$  operation. (see Section 14.5 below). If the left side of an  $=$  sign is an integer variable and right side has a real value, then the integer variable is given the integer value  $\leq$  the real variable on right. For example for

$$I = 5.71$$

$$J = 1.001$$

$$K = 0.999999$$

the computer assigns the integer values 5, 1 and  $0$  for  $I$ ,  $J$ , and  $K$  respectively. If we write

$$I = 5.71$$

$$C = I$$

then  $C$  will be assigned the value  $5.0$

#### 14.4 FORTRAN library functions

Some commonly used functions are already programmed and can be used by simply calling the appropriate names. Some of these are (for a complete list see Rajaraman or any other book on FORTRAN) :

$$\sqrt{X} \rightarrow \text{SQRT}(X)$$

$$|X| \rightarrow \text{ABS}(X)$$

$$\sin X \rightarrow \text{SIN}(X)$$

$$\cos X \rightarrow \text{COS}(X)$$

$$\sin^{-1} X \rightarrow \text{ASIN}(X)$$

$$\cos^{-1} X \rightarrow \text{ACOS}(X)$$

$$\tan^{-1} X \rightarrow \text{ATAN}(X)$$

$$\sinh X \rightarrow \text{SINH}(X)$$

$$\cosh X \rightarrow \text{COSH}(X)$$

$$e^X \rightarrow \text{EXP}(X)$$

$$\ln X \rightarrow \text{ALOG}(X)$$

$$\log_{10} X \rightarrow \text{ALOG10}(X)$$



The function FLOAT can be used to convert an integer quantity to a real quantity:

$I = 5$

$X = \text{FLOAT}(I)$

will assign the value 5.0 to the real variable X. Function IABS gives the absolute value of an integer quantity:

$I = -3$

$J = \text{IABS}(I)$

will cause  $J = 3$ .

#### 14.5 Mixed mode operation.

When integer quantities are used on both sides of an operand, the evaluated quantity becomes an integer and unless care is taken, it may give wrong results altogether. For example, suppose we wish to program the expression  $\frac{1}{2}X^2 + b^2$ .

If we write

$C = (1/2)*X**2 + B**2$

then while evaluating (1/2) the computer encounters integers on both sides of the operand / and the result will be an integer. Since the integer conversion of a number will give an integer  $\leq$  the number, (1/2) will equal 0 and the expression will be wrong. The correct way to program it will be

$C = (1./2.)*X**2 + B**2$

In this case both the constants on either side of / are real and the value will be 0.5. Care should be taken whenever integer division is used. This is also true with integer variables in mixed mode, for which FLOAT should be used as

$(4/I)*X**2$  must be replaced by  $(4./\text{FLOAT}(I))*X**2$

#### 14.6 Input output statements.

There are two ways in which input data (necessary for the calculation) may be supplied to the computer.

(1) If only a few numbers are needed they can be fed directly in the program. Ex.1

$A = 5.7$

$B = 2.79$

$C = -1.509$

$D = 3.91E-12$

$K = 3$

One can also use the DATA statement: the list of variables, separated by commas, is given following DATA and followed by the values of the variables, separated by commas, in the same sequence and enclosed between slashes. Ex.2 — the same set of data as above can be given as:

DATA A,B,C,D,K/5.7,2.79,-1.509,3.91E-12,3/

Values of subscripted variables can also be given using the DATA statement. Ex.3

— For  $a_1 = 1.5$ ,  $a_2 = 2.1$ ,  $a_3 = 5.0$ ,  $b_1 = 0.3$ ,  $b_2 = 0.7$ , we can have

DIMENSION A(3), B(2)

DATA A/1.5,2.1,5.0/,B/0.3,0.7/

(2) Data (particularly a large set of data) can be read using the READ statement.

Ex. 4 — Data of Ex. 1 can be read by

READ(\*,\*) A,B,C,D,K

The first \* within the parentheses means that the data will be supplied on the screen using key board *after* the command for execution is given as (separated by commas)

5.7,2.79,-1.509,3.91E-12,3 ⇐

This should be followed by pressing the "Enter" key (denoted by the symbol ⇐ here). The second \* in Ex. 4 indicates free format.

The data can also be supplied from disk files created earlier and declared in the FORTRAN program by an OPEN statement. For example, the data of Ex. 1 can be supplied in a file named INPUT.DAT created before execution command is given. Then program should contain

OPEN (5, FILE = 'INPUT.DAT')

READ (5,\*) A,B,C,D,K

Output of results can be obtained by using the WRITE statement. Ex. 5 — Suppose we want the calculated quantities X, Y, A, K to be written in the screen; then

WRITE (\*,\*) X, Y, A, K,

The stars have the same meaning as in the READ statement. If we wish to specify a particular format, we can use FORMAT statement as

WRITE(\*,10) X, Y, A,K

10 FORMAT(3E16.6,110)

Here 10 refers to the FORMAT statement which says that X, Y and A should be written in exponent form within a field of length 16 with six digits after the decimal point for each of the variables. The integer variable K will be written within a field of 10 spaces (with the last digit of the value of K written at the end of this specified field). If we write

10 FORMAT (3E16.6,5X,110)

Then five extra spaces will be given before the integer variable is written in 110 format.

One can also write the output in an output file named, say OUT.DAT defined by the OPEN statement as

OPEN(6,FILE= 'OUT.DAT')

WRITE(6,10) X, Y, A,K

10 FORMAT(3E16.6,110)



(For most PC's the file OUT.DAT must be created as a blank file *before* execution of the program).

Then nothing will be written on the screen, but after execution of the program, one can print the file named OUT.DAT (on which results are written automatically during execution of the program), containing the values of the variables X, Y, A, K. As the file OUT.DAT is stored permanently in the hard disk, it can be read later or transferred to another computer using a floppy disk.

If we wish to write a particular comment in the output, we can use the WRITE statement, with the intended comment between quotes ( ' ') as shown below:

WRITE (\*,\*) 'The values of X, Y, A and K are:' The exact expression within the quotes will then be written in the screen.

### 14.7 Examples of simple programs

Ex. 1 — Roots of a quadratic equation  $ax^2 + bx + c = 0$ , with  $a=5.7$ ,  $b=2.79$ ,  $c=-1.509$ .

We have  $x = [-b \pm \sqrt{(b^2-4ac)}]/2a$

$$A = 5.7$$

$$B = 2.79$$

$$C = -1.509$$

$$R = B*B-4.*A*C$$

$$R = \text{SQRT}(R)$$

$$X1 = (-B+R)/(2.*A)$$

$$X2 = (-B-R)/(2.*A)$$

WRITE(\*,\*) 'SOLUTIONS ARE → ', X1,X2

Ex.2 — Area and sides of a triangle whose two sides are a and b and the angle included is  $\theta$  (i.e. degrees). Given  $a = 4.5$ ,  $b = 2.73$ ,  $\theta = 40^\circ$ . The length of the third side is given by

$$C = \sqrt{(a^2 + b^2 - 2ab \cos \theta)}$$

$$\text{and area} = ab \sin \theta / 2$$

Suppose a,b, $\theta$  are given in a file named TRIAN.DAT in free format and we want results written in another file named OUTPUT.DAT

OPEN(5,FILE = 'TRIAN.DAT')

OPEN(6,FILE = 'OUTPUT.DAT')

READ(5,\*)A,B,THETA

$$\text{THETA} = \text{THETA} * 3.141593 / 180.0$$

$$C = \text{SQRT}(A**2+B**2-2.*A*B*\text{COS}(\text{THETA}))$$

$$\text{AREA} = A*B*\text{SIN}(\text{THETA})/2.0$$

```

WRITE(6,20)C, AREA
20  FORMAT(2X,'C=', E16.6,5X,'AREA=', E16.6)
STOP
END

```

Execution of this program will not produce any result on the screen, but values of C and area will be written in the file OUTPUT.DAT (This must be created as a blank file as shown in this example before executing the program). If we wish to see this on the screen, we can give the screen command

```
TYPE OUTPUT.DAT ⇐
```

If we wish to print the output on paper, we can give the screen command

```
PRINT OUTPUT.DAT ⇐
```

#### 14.8 Control statements.

It is possible to have conditional execution in the program using the IF statement. For ex. if  $b^2 \geq 4ac$ , then calculate  $\sqrt{b^2 - 4ac}$ . This can be done by

```

D = B**2
E = 4.*A*C
IF (D .GE. E) F = SQRT(D-E)

```

The expression (D .GE. E) has only two values — “true” (if  $D \geq E$ ) or “false” (if  $D < E$ ). If the expression is “true” then F is calculated, if “false” then F is not calculated. The operator .GE. is called a “relational operator”. Note that a dot (.) on either side is required. The relational operators in FORTRAN are:

.GT.	for	>
.GE.	for	≥
.LT.	for	<
.LE.	for	≤
.EQ.	for	=
.NE.	for	≠

Using the GO TO statement, one can have conditional branching. The statement GO TO 30

means that when this statement is reached in the normal sequence, the execution will go to statement no. 30 unconditionally. We can have conditional branching as shown below:

```

D = B**2
E = 4.*A*C
IF(D .GE. E) GO TO 30
F = 0.0

```



GO TO 50

30 F = SORT(D-E)

50 CONTINUE

In this case if  $D > E$ , execution will jump from 3rd to the 6th statement, otherwise F will be set equal to 0 and control will go to statement no 50.

Note the statement CONTINUE. This is an executable statement which does nothing. It merely asks the machine to continue execution. The main use of this statement is as a delimiter of a group of statements, especially in DO loops (see section 14.9).

We can also have comparison between computed numbers. For ex., the same results will be obtained by the sequence of statements:

F = 0.

IF ((B\*\*2).GE.(4.\*A\*C)) F = SQRT(B\*\*2-4.\*A\*C)

Here F is set equal to 0 first and then if  $b^2 \geq 4ac$ , then F is calculated as  $\sqrt{b^2-4ac}$ , otherwise F is not recalculated; hence it retains its previous value (=0).

It is possible to use the IF statement to cause branching in the following manner:

IF (A) 10,20,30

10 B = (C+D)/A

.

.

.

GO TO 100

20 B = C-D

.

.

.

GO TO 100

30 B = (C+D)/SQRT(A)

.

.

.

100 CONTINUE

The IF statement causes the execution to go to statement no. 10 if  $A < 0$ , to statement no. 20 if  $A=0$  and to statement no. 30 if  $A > 0$ .

Another form of branching can be obtained as in the following example:

GO TO (10,20,30,40) KOUNT

This causes transfer of control to statement nos. 10,20,30 or 40 (which must be executable FORTRAN statements and not, e.g., a FORMAT statement) if KOUNT has values 1,2,3,4 respectively. If KOUNT has a value < 1 or > the total number of FORTRAN statements within the parentheses, the control is transferred to the statement immediately following to GO TO statement.

### 14.9 The DO statement

If it is necessary to repeat a set of statements a fixed number of times, one can use the DO statement. It has the general form

```
DO 100 I = M1,M2,M3
```

```
.  
.  
.
```

```
100 CONTINUE
```

I must be an integer variable (i.e must not be defined otherwise) and M1,M2,M3 are integers (constants, variables or expressions). Then all the statements between DO and 100 CONTINUE are executed repeatedly with the value of  $I = M1, M1+M3, M1+2 \cdot M3, \dots$  so long as  $I \leq M2$ . The last statement need not be a CONTINUE statement, it may be any executable statement; however if we wish the execution to be skipped for a particular value of I, determined by a conditional GO TO statement within the DO loop, then a CONTINUE statement at the end of the loop is needed.

If M3 is not given its value is assumed to be 1. For ex.

```
DO 100 I = 5,10
```

means that the loop will be repeated for  $I = 5,6,7,8,9,10$ .

Ex. 1 : Suppose we have to calculate the sum

$$S = 1 + x + x^2 + x^3 + \dots + x^k$$

This can be programmed as

```
READ(*,*) X,K
```

```
S = 1.
```

```
T = 1.
```

```
DO 10 I = 1, K
```

```
T = T*X
```

```
S = S+T
```

```
10 CONTINUE
```

```
WRITE(*,*) 'S=', S
```

```
STOP
```

```
END
```



Ex. 2 : Calculate the factorials of integers upto 30!

DIMENSION FAC(31)

C FAC(I) = FACTORIAL OF (I-1)

FAC(1) = 1.

FAC(2) = 1.

DO 20 I = 3,31

20 FAC(I) = FAC(I-1)\*FLOAT(I-1)

WRITE(\*,\*) 'FACTORIALS ARE'

WRITE(\*,30) (FAC(I),I = 1,31)

30 FORMAT (5E16.6)

STOP

END

In the above example, the subscripted real variable FAC is used to store the factorials as  $FAC(I) = (I-1)!$ . This is mentioned in the second statement called a "comment statement" (such statements must begin with a C in the first column and is not processed by the computer – it is meant as a reminder to the program writer). In the above example, an implicit DO loop appears in the 8th statement, where FAC(1), FAC(2), ....., FAC(31) will be written according to 30 FORMAT (five values in one line).

It is possible to nest several DO loops one within another (called nested DO loops) as indicated schematically below:

DO 100 I = 1,10

.

.

DO 50 J = 1,5

.

.

50 CONTINUE

.

.

100 CONTINUE

Crossing of DO loops in the nesting is *not* allowed. For example:

```

DO 50 I = 1,10
.
.
DO 100 J = 1,5
.
.
50 CONTINUE
.
.
100 CONTINUE

```

is an *illegal nesting*. Jumping *into* a DO loop by conditional branching from outside is *not* allowed, although *jumping out of* a DO loop is possible.

#### 14.10 Numerical Methods

One can devise numerical methods (iterative or otherwise) to solve a given problem. We will solve only a commonly encountered numerical problem and its solution by a simple method. A large number of numerical formulas and useful tables can be obtained in "Handbook of Mathematical Functions" by M. Abramowitz and I. A. Stegun.

##### Solution of $f(x)=0$ by bisection method

Suppose we want to solve  $f(x)=0$ , where  $f(x)$  is a given function of  $x$ . If we know two values of  $x$  (say  $x_L$  and  $x_R$ ) at which  $f(x)$  has opposite signs and only one zero in between, we can adopt the following algorithm:

- (1) Read  $x_L$ ,  $x_R$  and  $e$  (precision needed).
- (2) Calculate  $f_L = f(x_L)$  and  $f_R = f(x_R)$
- (3) If  $f_L f_R < 0$  go to (5)
- (4) Otherwise write "wrong interval" and go to (1).
- (5) Calculate  $x_M = (x_L + x_R)/2$
- (6) If  $|x_L - x_R| \leq e$  go to (11)
- (7) Otherwise calculate  $f_M = f(x_M)$
- (8) If  $f_M f_L < 0$  go to (10)
- (9) Otherwise replace  $x_L$  by  $x_M$  and  $f_L$  by  $f_M$ , then go to (5)
- (10) Replace  $x_R$  by  $x_M$  and  $f_R$  by  $f_M$ , then go to (5)



(11) Write "solution is",  $x_M$

The program with comments is shown here for a function

$$f(x) = x^3 - 2.46x^2 + 2.5x - 4 = 0.$$

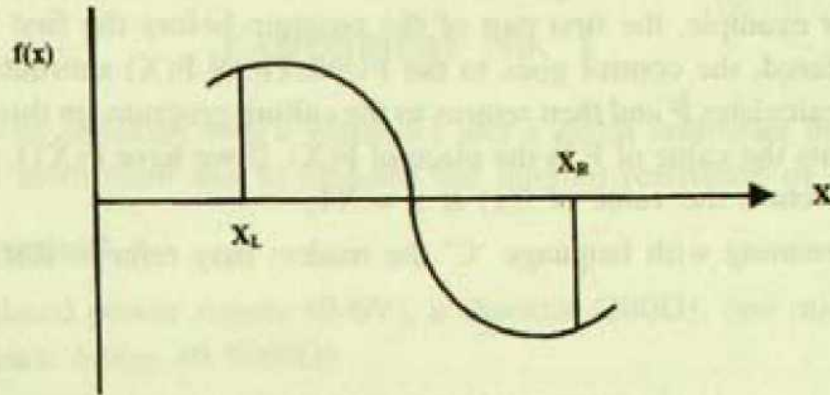


Fig. 14.1

```

C      PROGRAM TO SOLVE F(X) = 0
100  WRITE(*,*) 'SUPPLY XL,XR,E'
      READ(*,*) XL,XR,E
      FL = F(XL)
      FR = F(XR)
      IF(FL*FR .LT. 0.) GO TO 200
      WRITE(*,*) 'WRONG INTERVAL'
      GO TO 100
200  XM = (XL+XR)/2.0
      IF(ABS(XR-XL) .LE. E) GO TO 400
      FM = F(XM)
      IF(FM*FL .LT. 0.) GO TO 300
      XL = XM
      FL = FM
      GO TO 200
300  XR = XM
      FR = FM
      GO TO 200
400  WRITE(*,*) 'SOLUTION IS X =',XM
      STOP
      END
      FUNCTION F(X)
  
```

$F = X^{**3} - 2.46 * X^{**2} + 2.5 * X - 4.$

RETURN

END

In this case, we have an example of a FUNCTION subroutine. Whenever in the MAIN program (in our example, the first part of the program before the first END statement) F(X) is encountered, the control goes to the FUNCTION F(X) subroutine with the last value of X and calculates F and then returns to the calling program (in this case, the MAIN program) and puts the value of F in the place of F(X). If we have F(X1), the FUNCTION subroutine will return the value of f(x) at  $x = X1$ .

For programming with language 'C' the readers may refer to Ref. 3.

## References

1. Computer Programming in FORTRAN 77 — V. Rajaraman, 3rd edition (Prentice-Hall of India)
2. A Guide to FORTRAN IV Programming, McCracken, (John-Wiley)
3. Theory and Problems of Programming with C, B. S. Gottfried (Schaum Series - Tata McGraw Hill, 1998) Chapters 2-6



## Chapter 15

### Experiments

#### Experiment No. 1

To convert a given ammeter into a voltmeter and a given voltmeter into an ammeter. To calibrate the instrument and to measure the internal resistance of it in each case.

##### Instruments required :

Adjustable regulated power supply (0-6V), a rheostat (200Ω), one microammeter (0-100μA), resistance boxes (0-5000Ω)

##### I. To measure the internal resistance of the Microammeter

First set up the circuit as shown in the fig. 15.1

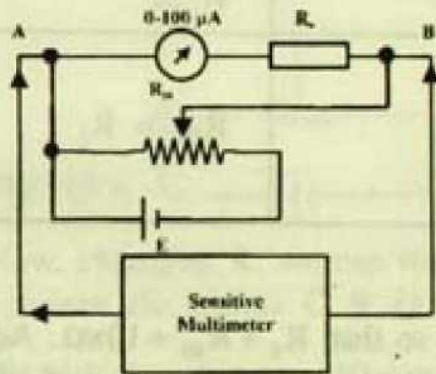


Fig. 15.1

If  $i$  is the current observed in the microammeter, then, the potential difference,  $(V_A - V_B) = i(R_m + R_s)$  where  $R_m$  is the internal resistance of the meter.

$(V_A - V_B)$  is measured by a sensitive multimeter for various  $i$  with a given  $R_s$  and a graph is drawn for  $(V_A - V_B)$  vs  $i$ .

The slope of the graph is  $R_s + R_m$ , knowing  $R_s$ ,  $R_m$  can be determined.

TABLE 1 (a) [ $R_s \geq \frac{E}{100\mu A}$  initially, otherwise the  $\mu A$  may be damaged.]

$R_s$ in kΩ	Reading in microammeter $i$ μA	Reading of the multimeter $(V_A - V_B)$ Volt	$R_m$

### Alternative half deflection method

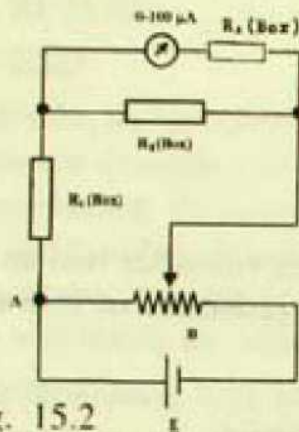


Fig. 15.2

Construct the circuit as shown in fig. 15.2. For a given resistance  $R_s$  (used as shunt) and  $R_2=0$ , adjust the resistance  $R_1$  so that the microammeter shows full scale deflection. Now insert resistances from the box  $R_2$  until the deflection in the microammeter is half (i.e.  $50\mu A$ ). The resistance  $R_2$  will then be the resistance of the microammeter. Repeat the experiment for different value of  $R_s$ .

TABLE 1 (b)

$R_s$ in $k\Omega$	$R_1$ in $\Omega$	$R_2$ in $\Omega$	Reading in the microammeter $\mu A$	Reading of the microammeter $R_m$ in $\Omega$
.....	.....	0	100	$R_m = R_2$
.....	.....	.....	.....	
		$R_2$	50	

### Conversion of microammeter to voltmeter

Remake the circuit (Fig. 15.1). Choose  $R_s$  so that  $R_s + R_m = 10k\Omega$ . Adjust the rheostat, observe the current  $i$  in the microammeter. The potential difference  $V_A - V_B = i(R_s + R_m)$ . Check that the p.d ( $V_A - V_B$ ) with a sensitive multimeter. Range of the meter is  $(100\mu A) \times (10k\Omega) = 1V$ . Repeat the procedure for  $(R_s + R_m) = 50k\Omega, 100k\Omega$  and so on and the  $100\mu A$  ammeter is converted to a 5V, 10V and so on voltmeter, respectively.

TABLE 2 : Microammeter used as a dc voltmeter.

No of observation	$R_s$ in $k\Omega$	$R_m$ in $\Omega$	$R_s + R_m$ in $\Omega$	Microammeter reading $i$ $\mu A$	$(V_A - V_B)$ $= i(R_s + R_m)$ volt (a)	Reading with multimeter (b)	Change (a-b) in volt.





## Experiment No. 2

To construct an adjustable voltage power supply using IC and to study its regulations.

### Instruments required :

0-12v, 15v, 18v, 1Amp transformer, bread board, IC LM 317 regulator, 4.7k $\Omega$  potentiometer, rheostat (200 $\Omega$ ), resistance box, voltmeter, milliammeter.

### Principle :

An ideal regulated power supply is an electronic circuit which provides a fixed output voltage that remains fixed or regulated over a wide range of load current or input voltage or temperature variation. Starting with an a-c supply, a steady d-c voltage can be developed by rectifying the a-c voltage, then filtering to a d-c level we get an unregulated supply. We can regulate this d-c voltage by using IC voltage regulators. A group of linear IC voltage regulators that provide fixed output voltage is available. A block diagram containing parts of typical power supply is shown.

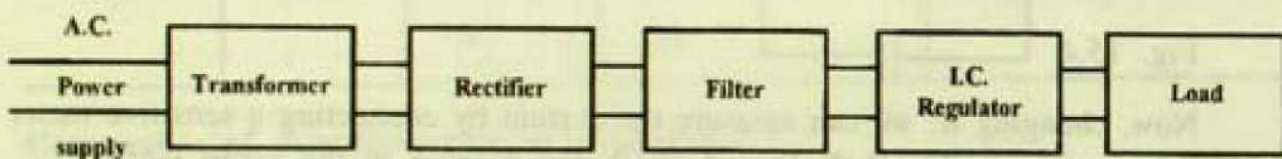


Fig. 15.5

### IC voltage regulators :

These regulators comprise a class of widely used ICs. These units contain circuitry for reference source, error amplifiers, control device and overload protection all in a single chip. There are popular 3-terminal voltage regulators providing positive, negative fixed output voltages or allowing an adjustable output voltage. One pin is for the unregulated input voltage, one pin is for the regulated output voltage and the one for the ground.

For a particular IC unit, device specification sheet provides the voltage range over which the input voltage can vary to maintain the regulated output voltage over a wide range of load current. An output-input voltage differential must be maintained for the IC to operate, which means that variation of input voltage must be kept large enough to maintain a voltage drop across the IC to permit proper operation of the internal circuit. The LM317 series needs an input voltage at least 2.5v greater than the regulated output voltage, otherwise it will stop regulating.



The actual circuit of operation for the regulated power supply using 3 pin IC regulator is given below.

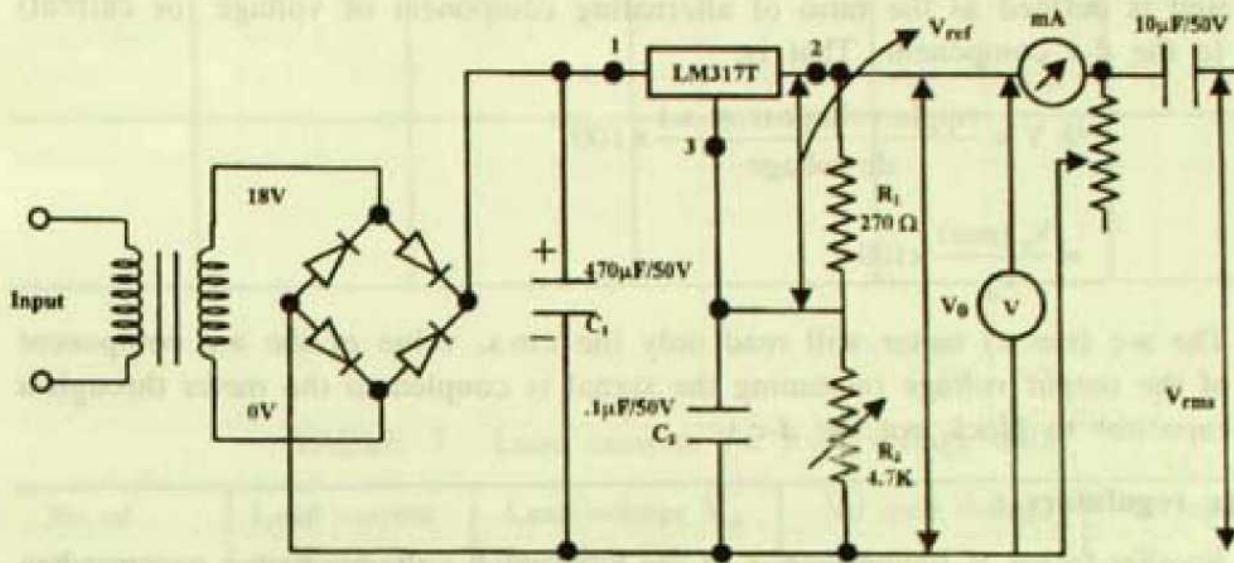


Fig. 15.6

Here the common terminal is not grounded but is connected to the top of  $R_2$ . This means the reference voltage is across  $R_1$ . A quiescent current  $I_Q$  flows through the pin 3 and through  $R_2$ . Therefore, the output voltage from pin 2 to the ground is

$$V_0 = V_{ref} + \left( I_Q + \frac{V_{ref}}{R_1} \right) \cdot R_2$$

Since,  $I_Q$  shows little variation with line and load changes,  $V_{out} \equiv V_0$  is regulated and adjustable. Selection of resistors  $R_1$  and  $R_2$  allows setting of the output to the desired voltage.

For LM317, typical value for  $V_{ref} = 1.25\text{V}$  and  $I_Q = 100\text{ }\mu\text{A}$ . Therefore, for  $R_1 = 270\Omega$  and  $R_2 = 2.4\text{ k}\Omega$ ,

$$\begin{aligned} V_0 &= 1.25\text{V} \left[ 1 + \frac{2.4\text{k}\Omega}{270\Omega} \right] + 100\mu\text{A} (2.4\text{k}\Omega) \\ &= 12.39\text{V} + .24\text{V} \\ &= 12.63\text{V} \end{aligned}$$

The capacitor  $C_2$  is used to avoid any unwanted oscillation due to lead inductance within IC.

## Ripple factor Y

It measures the smoothness of the d-c output of the regulated power supply and is defined as the ratio of alternating component of voltage (or current) to the d-c component. That is,

$$\begin{aligned} \% Y &= \frac{\text{ripple voltage (r.m.s.)}}{\text{dc voltage}} \times 100 \\ &= \frac{V_r (\text{rms})}{V_{dc}} \times 100 \end{aligned}$$

The a-c (r.m.s.) meter will read only the r.m.s. value of the a-c component of the output voltage (assuming the signal is coupled to the meter through a capacitor to block out the d-c.)

## Voltage regulators :

Another factor of importance is to see how much voltage change occurs when the output of a power supply (under no load condition) is connected to a load. This is described by a factor called voltage regulation. This is defined by

$$\% \text{ Voltage regulation} = \frac{V_{NL} - V_{RL}}{V_{RL}} \times 100$$

where  $V_{NL} \rightarrow$  Voltage at no load

$V_{RL} \rightarrow$  Voltage at rated load

## Input regulation factor ( $S_v$ )

It is defined as the ratio of the change in output voltage ( $\Delta V_o$ ) to the change in input voltage ( $\Delta V_i$ ) for constant load current.

## Procedure :

First perform the circuit as shown in the figure 15.6. Observe the load voltage at various load currents for a given unregulated voltage. Repeat the observation with a different unregulated voltage and so on. Record all the observations according to the following tables.

TABLE 1 : Specification of the meters used

Meter used for measurement	Range of the meter	Value of one division	Eye estimation	Error
Voltmeter for $V_o$				
Multiammeter for $I_o$				
Ac V/mv for ripple				



**TABLE 2 : Values of the circuit elements**


**TABLE 3 : Load current Vs. load voltage data**

No. of observation	Load current $I_L$ in mA	Load voltage $V_O$ in volt	No load voltage $V_{NL}$	% regulation $\frac{V_{NL} - V_{RL}}{V_{RL}} \times 100$

(A graph is drawn  $I_L$  Vs.  $V_L$  and regulation is calculated)

**TABLE 4 : For ripple factor Y**


The output will have the same polarity and almost the same magnitude as the input.

Also, when the OP-AMP is used in a voltage follower or current source load back circuit, its input impedance becomes very large.

TABLE 5 : For the input regulation factor

Load current in mA	Input voltage $V_i$	Change in input voltage $\Delta V_i$	Output voltage $V_o$	Change in output voltage $\Delta V_o$	$S_v = \frac{\Delta V_o}{\Delta V_i}$

Note : To find the input regulation factor  $S_v$ , the transformer should be chosen so that the secondary has various tapings 12V, 15V, 18V. For continuous variation, one can use a variac at the primary of the input transformer.

To study the regulation characteristic, the students may be advised to keep the load voltage 10V-15V and the maximum load current 100 to 200 mA for longer durability of the apparatus.



## Experiment No. 3

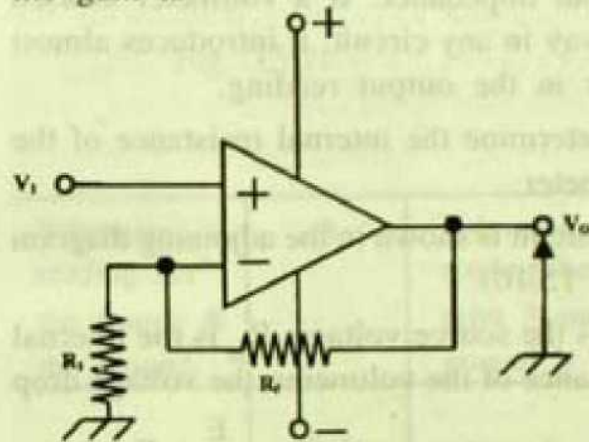
To measure the internal resistance of an analog voltmeter and to increase its internal resistance by using an OPAMP.

### Instruments :

+12v – 0 – 12V regulated power supply, one IC741, a voltmeter (0-10v), one adjustable regulated power supply and resistances.

### Principle :

When the OP-AMP is used as a noninverting amplifier (as shown in Fig. 15.7), the gain is



$$\frac{V_o}{V_i} = 1 + \frac{R_f}{R_1} \quad \text{If } R_f = 0$$

then  $\frac{V_o}{V_i} = 1$ , i.e.  $V_o = V_i$ , the circuit becomes that of the voltage follower (unity follower) (Fig. 15.8)

Fig. 15.7

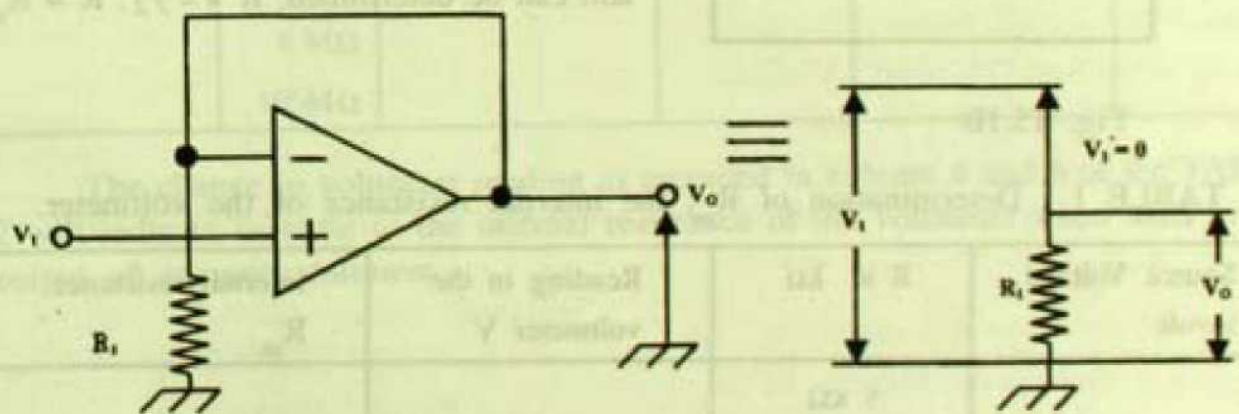


Fig. 15.8

The output will have the same polarity and almost the same magnitude as the input.

Also, when the OP-AMP is used in a voltage series or current series feed back circuit, its input impedance becomes very large.

If the open loop input impedance of an OP-AMP is  $r_{in}$ , then its closed loop input impedance is

$r_{in}(CL) = r_{in}(1+AB)$ , where  $A$  is the internal gain of the OP-AMP and  $B$  is the feed back factor. In the case of a voltage follower,  $B = 1$ . Therefore,  $r_{in}(CL) = r_{in}(1+A) = A r_{in}$  (Since  $A$  is very large). For an OPAMP 741C, typical values are :  $A = 100,000$  and,  $r_{in} = 2M\Omega$ , so  $r_{in}(CL) = 2 \times 10^{11} \Omega$

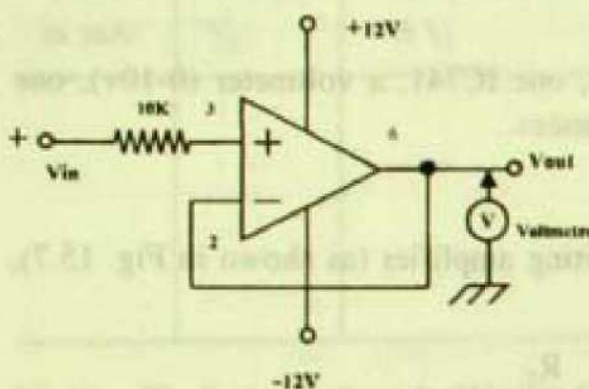


Fig. 15.9

Hence, if a voltmeter is connected as shown in the circuit (Fig. 15.9), it would have very high input impedance. If a voltmeter is used in this way in any circuit, it introduces almost no error in the output reading.

1. To determine the internal resistance of the voltmeter.

The circuit is shown in the adjoining diagram (Fig. 15.10).

If  $E$  is the source voltage,  $R_m$  is the internal resistance of the voltmeter, the voltage drop across the voltmeter is  $V = \frac{E}{R + R_m} R_m$ . Now varying  $R$  and observing  $V$  for a given  $E$ ,  $R_m$  can be determined. If  $V = \frac{E}{2}$ ,  $R = R_m$ .

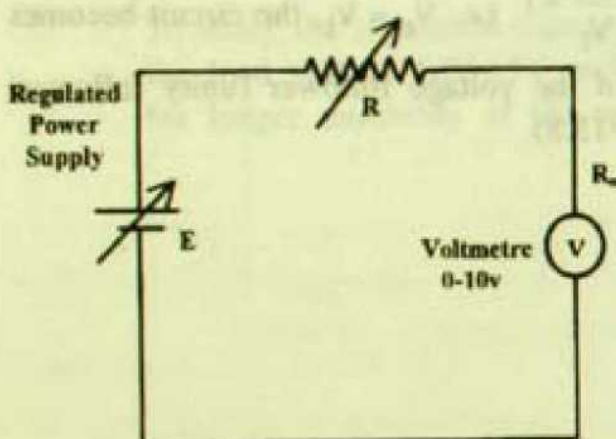


Fig. 15.10

TABLE 1 : Determination of  $R_m$ , the internal resistance of the voltmeter.

Source Voltage in volt	$R$ in $k\Omega$	Reading in the voltmeter $V$	Internal resistance $R_m$
	5 $k\Omega$		
	10 "		
	15 "		
	20 "		
	30 "		
	40 "		



II. To increase the internal resistance of the voltmeter construct the circuit as in (Fig. 15.11).

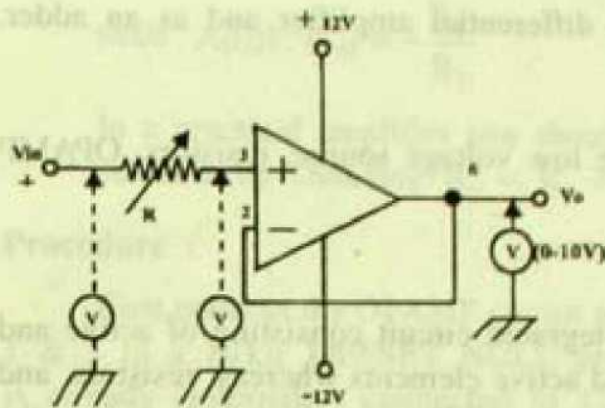


Fig. 15.11

First choose  $R = 10\text{ k}\Omega$ , take the voltmeter reading between the source and the ground, the point 3 and the ground and the point 6 and the ground and repeat the same for various  $R$ .

TABLE 2

Voltmeter reading bet <sup>n</sup> the source & the ground	R	Voltmeter reading bet <sup>n</sup> print 3 and the ground V	Change in $(V_{in} - v)$ volt	Voltmeter reading bet <sup>n</sup> in pt 6 and the ground $V_o$	Change in $(V_{in} - V_o)$
$V_{in}$	10 k $\Omega$			$V_o$	
	50 k $\Omega$				
	100 k $\Omega$				
	1 M $\Omega$				
	10 M $\Omega$				

The change in voltmeter reading as recorded in column 4 and 6 of the TABLE 2 will indicate increase in the internal resistance of the voltmeter when used in the output of an unity follower.

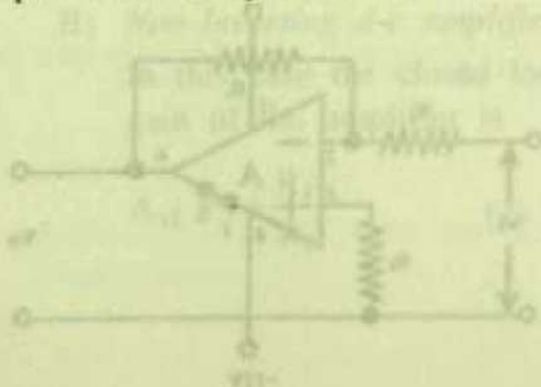


Fig. 15.12

## Experiment No. 4

To use OP-AMP as inverting, noninverting, differential amplifier and as an adder.

### Apparatus required :

A regulated +12V-0 – 12V supply, a variable low voltage source, resistors, OPAMP (741C), voltmeter, switches.

### Principle :

This OPAMP is a monolithic silicon integrated circuit consisting of active and passive elements. Transistors, diodes are called active elements whereas, resistors and capacitors are called passive elements. It is not possible to produce inductor on the chip by diffusion technique. The basic OPAMP is a direct coupled high gain amplifier consisting of one or more differential amplifier, followed by an emitter follower, level shifter and the output stage. The output is controlled externally by the addition of negative feedback. The OPAMP was originally designed to perform various mathematical operations. It is a versatile device and is being used in thousand of diverse applications.

The symbol used to represent the OPAMP is shown in the figure. The OPAMP has two inputs and one output. One input is called the inverting input and is denoted by (–) sign. When a signal is applied at this input it will suffer a phase change of  $180^\circ$  at the output. The second terminal is called the noninverting (+) input. A signal applied to this input will appear as an amplified output with the same phase.

The amplifier has a stable voltage gain, high input impedance and low output impedance. Some of its uses are given below.

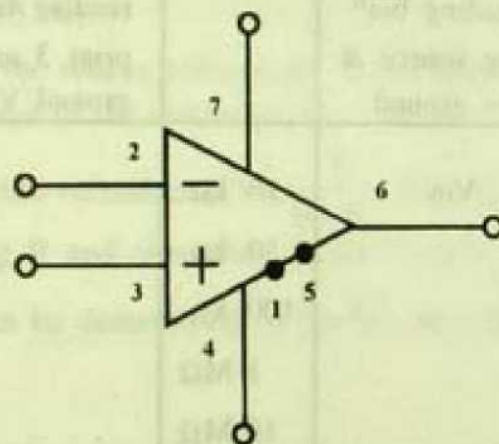


Fig. 15.12

#### A) Inverting d-c amplifier

The circuit diagram of an inverting d-c amplifier is shown in fig. 15.13.

The closed loop gain of the amplifier is given by

$$A_{vf} = \frac{V_o}{V_i} = -\frac{R_f}{R_1} \left( 1 - \frac{1}{1 + A\beta} \right)$$

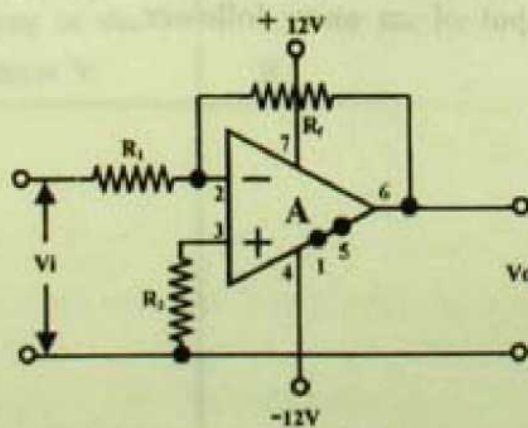


Fig. 15.13



where  $A$  is the open loop gain and  $\beta$  is the feed back fraction  $= \frac{R_1}{R_1 + R_f}$

since  $A\beta \gg 1$ ,  $A_{vf} = -\frac{R_f}{R_1}$  (1)

In a practical amplifier one should minimise the offset error due to input bias currents by choosing  $R_2 = R_1 \parallel R_f$

### Procedure :

First connect the OPAMP circuit as shown (Fig. 15.13) and connect the terminals 1 & 5 to a 10k $\Omega$  OFFSET NULL potentiometer. The centre of the potentiometer is already (internally) connected to 12V terminal. Adjust the potentiometer to get zero output voltage. Now, observe the output for various input at the inverting terminal. Repeat the experiment with a different gain.

### Results :

TABLE 1

Input voltage $V_i$	$R_1$	$R_f$	$V_{out}$	$A_v = \frac{V_{out}}{V_i}$	Theoretical gain

### B) Non-Inverting d-c amplifier

In this case the closed loop gain of the amplifier is

$$A_{vf} = 1 + \frac{R_f}{R_1} \quad (2)$$

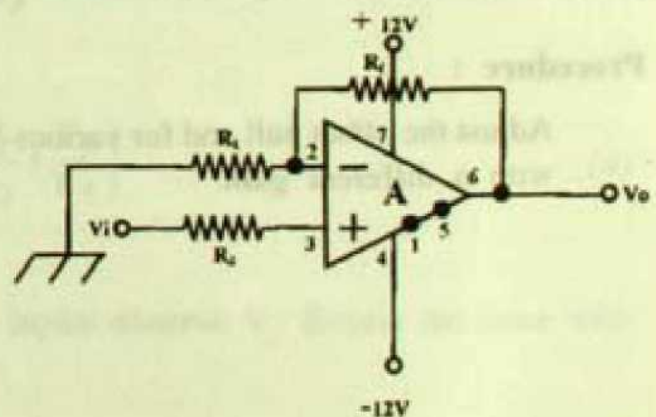


Fig. 15.14

**Procedure :**

Adjust the offset null and proceed as before.

**Results :**

TABLE 2

Input voltage $V_i$	$R_i$	$R_f$	Output voltage $V_o$	Gain $A_v = \frac{V_o}{V_i}$	Theoretical gain $= 1 + \frac{R_f}{R_i}$

**C) The differential amplifier**

In this case,  
Output voltage,

$$V_o = \frac{R_f}{R_i} \Delta V$$

where  $\Delta V = V_2 - V_1$

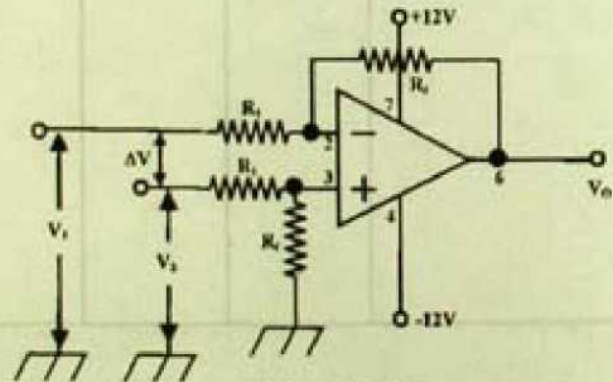


Fig. 15.15

**Procedure :**

Adjust the offset null and for various  $V_1$  &  $V_2$ , Observe  $V_o$ . Repeat the experiment with a different gain.



**Results :**

**TABLE 3**

$R_1$ in $\Omega$	$R_f$ in $\Omega$	$V_1$ in volts	$V_2$ in volts	$V_3$ in volts	$V_o$ in volts Experimental	$V_o$ in volts Calculated

**D) Adder**

The meeting point of the three resistors is virtually at the ground potential, so the currents  $i_1, i_2, i_3, \dots$  will flow through respective input resistors uninfluenced by each other i.e.

$$i_1 = \frac{V_1}{R_1}, \quad i_2 = \frac{V_2}{R_2}, \quad i_3 = \frac{V_3}{R_3}$$

and  $i = i_1 + i_2 + i_3$

since  $V_o = -R_f i = -R_f \left( \frac{V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_3}{R_3} \right)$  (4)

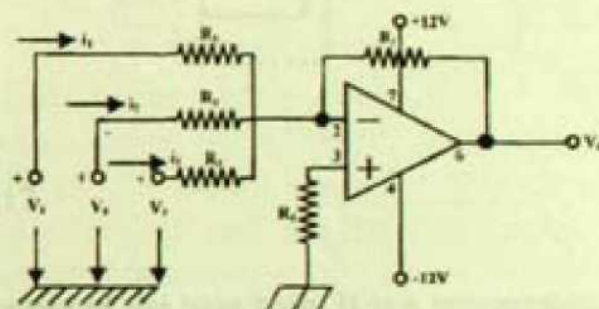


Fig. 15.16

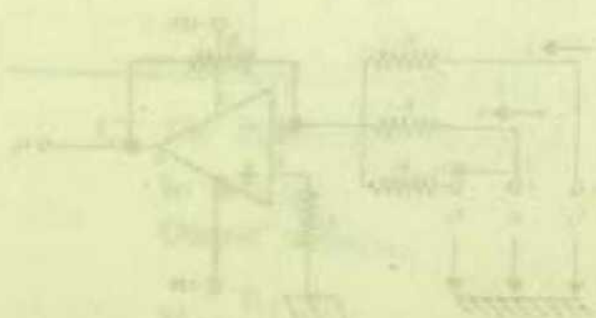
**Procedure :**

Adjust the offset null and for various inputs observe  $V_o$ . Repeat the same with a different  $R_f$ .

**Results :**

**TABLE 4**

Resistances $R_1, R_2, R_3$	$R_1$	$V_1$ in volts	$V_2$ in volts	$V_3$ in volts	Output $V_o$	Theoretical output
$R_1 = R_2 = R_3$ $= 10 \text{ k}\Omega$	10 k $\Omega$					
	100 k $\Omega$					



The inverting point of the op-amp is at virtual ground. Hence, the current  $i_1$  through  $R_1$  is  $i_1 = \frac{V_1}{R_1}$ . Similarly, the current  $i_2$  through  $R_2$  is  $i_2 = \frac{V_2}{R_2}$  and the current  $i_3$  through  $R_3$  is  $i_3 = \frac{V_3}{R_3}$ . The output voltage  $V_o$  is given by  $V_o = -R_f (i_1 + i_2 + i_3)$ .

$$V_o = -R_f \left( \frac{V_1}{R_1} + \frac{V_2}{R_2} + \frac{V_3}{R_3} \right)$$

**Procedure :**

- (1) Adjust the offset null and for various inputs observe  $V_o$ . Repeat the same with a different  $R_f$ .
- (2) With  $V_1 = 10 \text{ V}$ ,  $V_2 = 0 \text{ V}$ ,  $V_3 = 0 \text{ V}$ , observe  $V_o$  for different values of  $R_1$ .

**Procedure :**

- (1) Adjust the offset null and for various inputs observe  $V_o$ . Repeat the same with a different  $R_f$ .
- (2) With  $V_1 = 10 \text{ V}$ ,  $V_2 = 0 \text{ V}$ ,  $V_3 = 0 \text{ V}$ , observe  $V_o$  for different values of  $R_1$ .







will act as a trigger which drives the gate of an SCR (Silicon Controlled Rectifier) which latches and remains closed so long as the output of the comparator (OPAMP2) remains positive. The output of the comparator should supply necessary firing potential of the SCR. In the above circuit TYN 6004 SCR which can handle current of 4 Amp requires a firing potential  $\sim 2V$ . To ensure the positive supply to the gate of an SCR an IN-4007 is used in the output of the comparator. With the SCR on, the heater placed in the water bath will be connected to the main. The thermistor within a metallic capsule is immersed along with the heating coil in the waterbath. With the increase of temperature of the waterbath the resistance of the thermistor will decrease and so the potential drop across it will also decrease. When the potential of the pin 3 of the comparator goes below that of the pin 2, the output will swing to the negative saturation and the SCR will be opened. Consequently, the heater will be turned off. Thus, when a certain temperature of the heat bath is reached, the heater will be turned on and off, thereby, maintaining a constant temperature of the bath.

### Procedure :

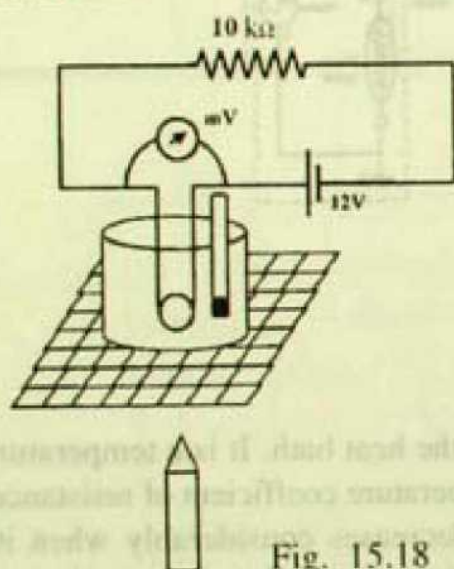


Fig. 15.18

First, place the thermistor encapsulated in a metallic cover within the water in a beaker which is heated with a burner. By inserting a thermometer within the beaker, the temperature of the bath and corresponding potential drop across the thermistor is observed at regular intervals. Draw a temperature Vs. potential drop across the thermistor curve. The calibration graph will enable one to select the reference potential of the comparator.

TABLE 1 : Reading for the potential drop across the thermistor and the temperature.

No of observation	Temperature of the heat bath	Mean temperature	Potential drop across the thermistor
1)	1) i) ii)	1)	
2)	2) i) ii)	2)	



**TABLE 2 : Setting of the reference voltage of the comparator and observation of the temperature of the heat bath.**

Reference voltage of the comparator volt	Corresponding temp. of the heat bath $O_c$ .	The temperature according to the calibration curve $O_c$
1.		
2.		
3.		

In making of the circuit (Fig. 15.17), first the live and neutral terminal of the ac main must be detected by an electrical tester. The neutral terminal must be connected to the ground while the live terminal should be connected to the anode of the SCR. This is necessary to avoid the possibility of having an electric shock while working with the ac mains. Alternatively, one can use relay instead of an SCR to turn the heater on and off. In that case, a transistor is necessary to activate the relay. The base of the comparator through a resistor ( $R_7$ ) while the relay is connected to the collector circuit (Fig. 15.19). From the point of view of safety the use of relay is advantageous.

**Note :** There are many possible ways, in which, the temperature controller circuit may be devised. The above procedure is adopted just to enable the students to have an idea about the application of an OP-AMP as comparator, the use of silicon controlled rectifier (SCR) as a switch and also, if one desires, the use of relay. The values of different resistors used in the circuit depend on the resistance of the thermistor at room temperature and the necessary trigger current and firing potential of the SCR. Choosing proper resistors, one can control the gain of the noninverting amplifier (OP-AMP 1) to a desired value and also by choosing the resistor  $R_7$  one can supply the necessary firing potential of an SCR.

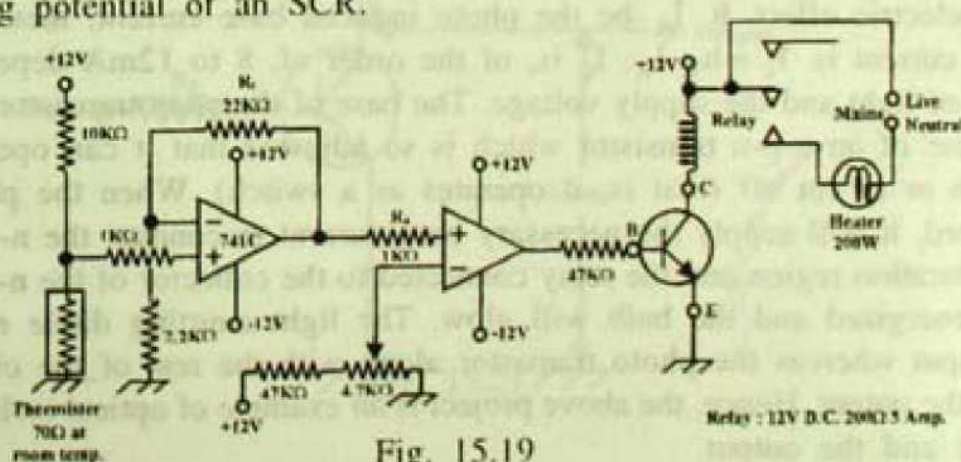


Fig. 15.19



## Experiment No. 6

To develop a photosensor using a photo transistor followed by an amplifier and to use the same to control the switching of a bulb.

### Components :

A +12V-0 – 12V regulated power supply, a photo transistor, a CL 100 transistor, a 12V DC 200Ω single contact 5Amp relay, a bulb (25W-100W).

The circuit diagram for the project is shown in the adjoining diagram (Fig. 15.20).

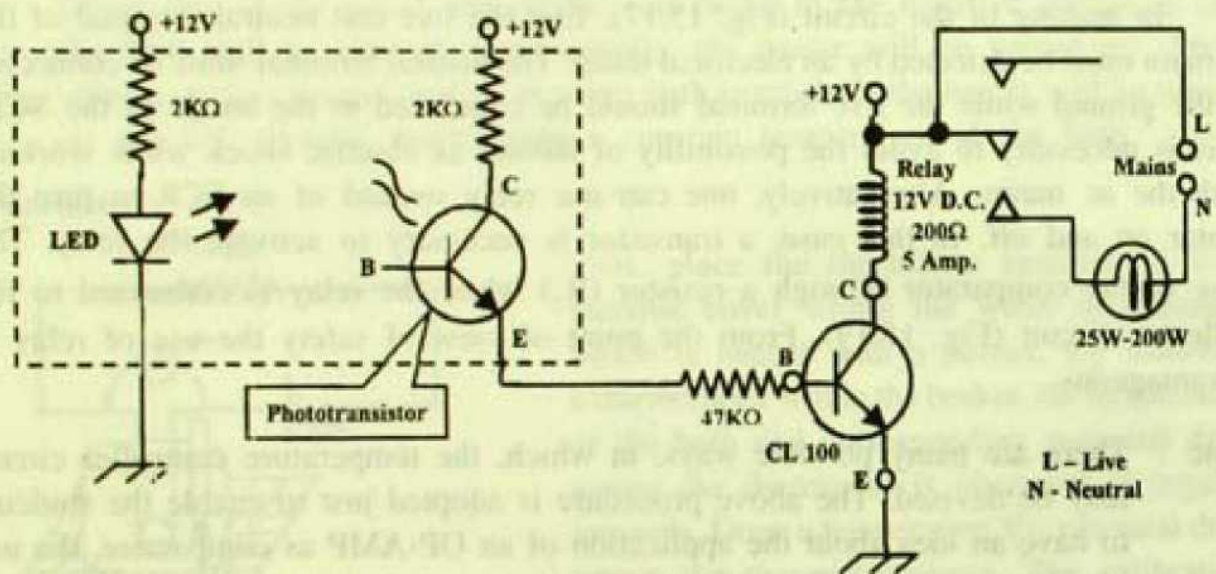


Fig. 15.20

### Principle :

The light emitting diode (LED) will emit light at an intensity determined by the forward current through LED. This light will activate the photo transistor, which has a photosensitive collector-base p-n junction. The light will supply base current by photoelectric effect. If  $I_\lambda$  be the photo induced base current, then the resulting collector current is  $I_c = h_{fe} I_\lambda$ .  $I_c$  is, of the order of, 8 to 12mA depending on the intensity of light and the supply voltage. The base of the phototransistor is connected to the base of an n-p-n transistor which is so adjusted that it can operate either at saturation or at cut off (that is, it operates as a switch). When the phototransistor is activated, it will supply the necessary base current to conduct the n-p-n transistor in the saturation region and the relay connected to the collector of the n-p-n transistor will be energized and the bulb will glow. The light emitting diode may be taken as the input whereas the photo transistor along with the rest of the circuit may be taken as the output. Hence, the above project is an example of optical isolation between the input and the output.



### Method :

To operate the n-p-n transistor either at saturation or cut off.

First, calculate the minimum base current needed to conduct the transistor in the saturation region as follows :

Let the resistor in the collector circuit be  $R_c$ . The supply voltage is  $V_{cc}$ . Therefore, Saturation value of the collector current is

$$I_c^{sat} = \frac{V_{cc} - V_{CE}^{sat}}{R_c} = \frac{V_{cc}}{R_c} \quad (V_{CE}^{sat} = 0.2V) \quad (1)$$

If  $\beta$  is the current gain of the transistor in the CE mode, the minimum base current needed to keep the transistor in the saturation region is  $I_B^{min} = \frac{I_c^{sat}}{\beta}$

The required base resistance is

$$R_B = \frac{V_{cc} - V_{BE}}{I_B^{min}} \quad (2)$$

For hard saturation, choose  $R_B \text{ (Hard)} \geq \frac{R_B}{10}$

To operate the transistor as a switch first put the transistor on a bread board and connect the leads to the supply through the resistors as shown (Fig. 15.21) and make the following observation.

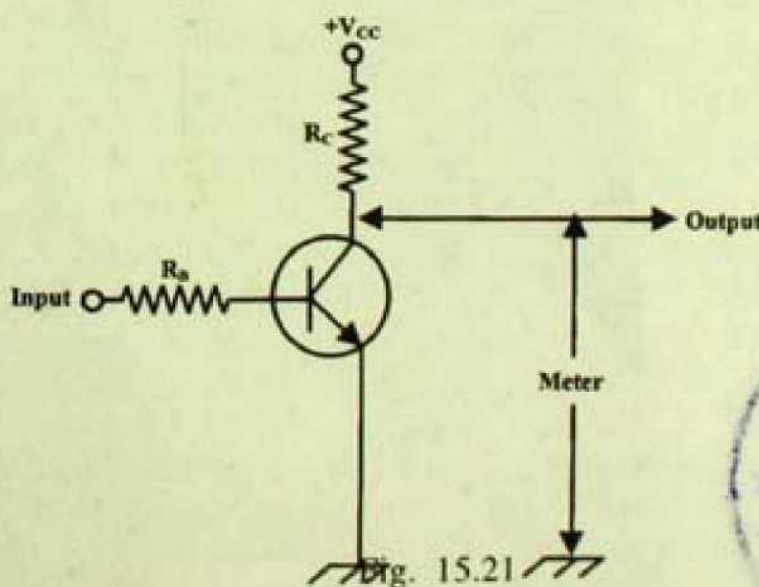


TABLE 1

$V_{cc}$ Volt	$R_c$	$R_B$	Input	Output	Inference
			0 +12V		

Note : The LED along with the phototransistor is to be kept under black cover. The phototransistor is sensitive to the sunlight. So, even without the use of LED and covering the phototransistor simply by hand one can demonstrate the project smoothly.

If one finds difficulty in using the relay one can perform the project by using a LED with a resistor 2K in the collector circuit instead of the relay.

